N° Ordre : ……/ Faculté des Sciences/UMBB/2016

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE M'HAMED BOUGARA - BOUMERDES

Faculté des Sciences

Thèse de Doctorat

Présentée par

M'hamed MATAOUI

Recherche d'information dans les documents XML: Prise en compte des liens pour la sélection d'éléments pertinents

Devant le jury:

Mohamed Ahmed-Nacer	Professeur à l'USTHB	Président
Amar Balla	Professeur à l'ESI	Examinateur
Rachid Ahmed-Ouamer	Professeur à l'UMMTO	Examinateur
Rabah Imache	Maître de conférences "A" àl'UMBB	Examinateur
Mohamed Mezghiche	Professeur à l'UMBB	Directeur de thèse
Mohand Boughanem	Professeur à l'IRIT (France)	Co-directeur de thèse
Tayeb Kenaza	Maître de conférences "A" à l'EMP	Invité

Année Universitaire :2015/2016

Order N°: ……/ Faculty of Sciences/UMBB/2015

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

M'HAMED BOUGARA UNIVERSITY OF BOUMERDES

Faculty of Sciences

Doctoral Thesis

Presented by

M'hamed MATAOUI

XML Information Retrieval:
Taking account of links for the selection of relevant elements

Jury members:

Mohamed Ahmed-Nacer	Full Professor at USTHB	President
Amar Balla	Full Professor at ESI	Examiner
Rachid Ahmed-Ouamer	Full Professor at UMMTO	Examiner
Rabah Imache	Senior Lecturer at UMBB	Examiner
Mohamed Mezghiche	Full Professor at UMBB	Thesis Advisor
Mohand Boughanem	Full Professor at IRIT	Thesis co-Advisor
Tayeb Kenaza	Senior Lecturer at EMP	Invited Member

Academic Year: 2015/2016

Résumé

Notre travail se situe dans le contexte de la *recherche d'information (RI)*, plus particulièrement la recherche d'information dans des documents semi structurés de type XML. L'exploitation efficace des documents XML disponibles doit prendre en compte la dimension structurelle. Cette dimension a conduit à l'émergence de nouveaux défis dans le domaine de la RI. Contrairement aux approches classiques de RI qui mettent l'accent sur la recherche des contenus non structurés, la RI XML combine à la fois des informations textuelles et structurelles pour effectuer différentes tâches de recherche. Plusieurs approches exploitant les types d'évidence ont été proposées et sont principalement basées sur les modèles classiques de RI, adaptées à des documents XML. La structure XML a été utilisée pour fournir un accès ciblé aux documents, en retournant des composants de document (par exemple, sections, paragraphes, etc.), au lieu de retourner tout un document en réponse une requête de l'utilisateur.

En RI traditionnelle, la mesure de similarité est généralement basée sur l'information textuelle. Elle permetle classement des documents en fonction de leur degré de pertinence en utilisant des mesures comme:« *similitude terme* » ou « *probabilité terme* ». Cependant, d'autres sources d'évidence peuvent être considérées pour rechercher des informations pertinentes dans les documents. Par exemple, les liens hypertextes ont été largement exploités dans le cadre de la RI sur le Web.Malgré leur popularité dans le contexte du Web, peud'approchesexploitant cette source d'évidence ont été proposées dans le contexte de la RI XML.

Le but de notre travail est de proposer des approches pour l'utilisation de liens comme une source d'évidencedans le cadre de la recherche d'information XML. Cette thèse vise à apporter des réponses aux questions de recherche suivantes :

1. Peut-on considérer les liens comme une source d'évidence dans le contexte de la RIXML?

2. Est-ce que l'utilisation de certains algorithmes d'analyse de liensdans le contexte de la RI XML améliore la qualité des résultats, en particulier dans le cas de la collection Wikipedia?

3. Quels types de liens peuvent être utilisés pour améliorer le mieux la pertinence des résultats de recherche?

4. Comment calculer le score lien des différents éléments retournés comme résultats de recherche? Doit-on considérer lesliens de type «document-document» ou plus précisément les liens de type «élément-élément»? Quel est le poids des liens de navigation par rapport aux liens hiérarchiques?

5. Quel est l'impact d'utilisation de liens dans le contexte global ou local?

6. Comment intégrer le score lien dans le calcul du score final des éléments XML retournés?

7. Quel est l'impact de la qualité des premiers résultats sur le comportement des formules proposées?

Pour répondre à ces questions, nous avons mené une étude statistique, sur les résultats de recherche retournés par le système de recherche d'information"*DALIAN*", qui a clairement montré que les liens représentent un signe de pertinence des éléments dans le contexte de la RI XML, et cecien utilisant la collection de test fournie par INEX. Aussi, nous avons implémenté trois algorithmes d'analyse des liens (Pagerank, HITS et SALSA) qui nous ont permis de réaliser une étude comparative montrant que les approches «*query-dependent*» sont les meilleures par rapport aux approches «*global context*» . Nous avons proposé durant cette thèse trois formules de calcul du score lien: Le premièreest appelée «*Topical Pagerank*»; la seconde est la formule : "*distance-based*"; et la troisième est :"*weighted links based*". Nous avons proposé aussi trois formules de combinaison, à savoir, la formule *linéaire*, la formule *Dempster-Shafer* et la formule *fuzzy-based*. Enfin, nous avons mené une série d'expérimentations. Toutes ces expérimentations ont montré que: les approches proposées ont permis d'améliorer la pertinence des résultats pour les différentes configurations testées; les approches *"query-dependent"* sont les meilleurescomparées aux approches *global context*; les approches exploitant les liens de type «élément-élément»ont obtenu de bons résultats; les formules de combinaison qui se basent sur le principe de l'incertitude pour le calcul des scores finaux des éléments XML permettent de réaliser de bonnes performances.

Mots-clés: recherche d'information, XML, Distance approche basée, approche de lien pondérée, analyse des liens, INEX, combinaison floue basée, combinaison de Dempster-Shafer.

Abstract.

Our work is in the context of information retrieval, particularly information retrieval (IR) in semi-structured XML documents. The effective exploitation of XML documents available requires to take the structural dimension into account. This dimension has led to the emergence of a new challenges in IR. Contrary to classical IR approaches that focus on searching unstructured content, XML IR combines both textual and structural information to perform different IR tasks. Several approaches exploiting the two types of evidences have been proposed and are mainly based on classic IR models adapted to XML documents. The XML structure has been used to provide a focused access to documents, by returning document component (e.g. sections, paragraphs, etc.), instead of whole document in response a user query.

In traditional IR, evidence of relevance is typically mined from textual evidence. It is based on ranking documents according to their degree of relevance using measures like: term similarity or term occurrence probability. However, other sources of evidence can be incorporated to retrieve relevant information in documents. For Instance, hyperlinks have been widely exploited in the context of Web retrieval. Despite their popularity in Web context, only few approaches, exploiting this source of evidence, have been conducted in the context of XML retrieval.

The aim of our work is to propose approaches for the use of links as a source of evidence in the context of XML information retrieval. This thesis aims to provide answers to the following research questions.

1. Can we consider the links as a source of evidence in the context of XML IR?
2. Does the use of some well-known link analysis algorithms in the XML Information Retrieval context improves the quality of results, particularly in the case of the Wikipedia collection?
3. What types of links can be used to best improve the relevance of retrieval results?
4. How to compute the link score of the different elements returned as retrieval results? Should we consider the "document-document" or more precisely "element-element" links? What is the weight the navigational links compared to hierarchical links?
5. What is the impact of using links in the global or local context?
6. How to incorporate the link score in calculating the final score of the returned XML elements?
7. What is the impact of the initial results quality on the behaviour of the proposed formulas?

Our proposals attempt essentially to address these questions:

We conducted a statistical study on the retrieval results returned by DALIAN IR system that showed clearly that the links represent a sign of relevance of the elements in the context of XML using the IR test collection provided by INEX. Also, we implemented three link analysis algorithms which allows us to perform a comparative study showing that *query-dependent* approaches are the best compared to global context approaches. We proposed during this thesis three link score computation formulas: The first is called"*Topical Pagerank*"; the second is the "*distance-based*"; and the third is the "*weighted links based*". We proposed three combination formulas, i.e. linear formula, Dempster-Shafer formula and fuzzy-based formula. Finally, we conducted a set of experiments including different parameters. All these experiments have shown that: the proposed approaches have improved the retrieval accuracy for the different tested configurations; the "*topic-sensitive*" approaches are the best compared to global context approaches; approaches exploiting the links of type "*element-element*" obtained good results; the combination formulas taking into account the uncertain aspects of computed scores of XML elements allow achieving good performance.

Keywords: information retrieval, XML, Distance based approach, weighted link approach, link analysis, INEX, fuzzy-based combination, Dempster-Shafer combination.

ملخص:

عملنا هذا يندرج في سياق البحث عن المعلومات، وخصوصا شبه منظمة. الاستغلال الفعال لملفات XML المتاحة يتطلب أخذ البعد الهيكلي في عين الاعتبار. وقد أدى هذا البعد إلى ظهور تحديات جديدة في هذا المجال. وخلافا للطرق الكلاسيكية التي تركز على البحث في محتوى غير منظم، فإن البحث عن المعلوماتفي ملفات XML يجمع بين المعلومات النصية والهيكلية لأداء المهامالبحث المختلفة. تم اقتراح عدة طرقلاستغلال هذين النوعين من المعلومات وتقوم أساسا على نماذج البحث عن المعلومات الكلاسيكية مكيفة لمعالجةملفات XML . تم استخدام هيكلية XML لتوفير وصول دقيق إلى الوثائق، من خلال العثور على مكوناتالملفات (على سبيل المثال أقسام، الفقرات، الخ)، بدلا من الملف بأكمله للإستجابة لإستعلام المستخدم.

فيالبحث عن المعلوماتالتقليدي، عادة ما يتم البحث عن طريق إستخدام الأدلة النصية. أي أنه يقوم بترتيب الملفات وفقا لدرجة الأهمية و ذلك باستخدام تدابير مثل: تشابه أو إحتمال تواجد مصطلح. ومع ذلك، يمكن أن تدمج مصادر أخرى من الأدلة للبحث عن المعلومات ذات الصلة في الوثائق. على سبيل المثال، تم استغلال الروابط التشعبية على نطاق واسع في سياق البحث في الويب. على الرغم من شعبيتها في سياق الويب، لا يوجد إلا القليل من الطرق فقط التي تقوم بإستغلال هذا المصدر من الأدلة، فإن البحث عن المعلوماتفي ملفات XML .

إن الهدف من عملنا هذا هو اقتراح طرق جديدة و مبتكرة لاستخدام الروابط كمصدر للأدلة في سياق البحث عن المعلوماتفي ملفات XML.وتهدف هذه الأطروحة إلى تقديم إجابات على الأسئلة البحثية التالية:

1. هل يمكننا إعتبار الروابط كمصدر للأدلة في سياقالبحثعنالمعلوماتفيملفاتXML؟
2. هلاستخدامبعضالخوارزمياتالمعروفةلتحليل الروابط فيسياقالبحثعنالمعلوماتفيملفاتXMLيحسننوعيةالنتائج،وخصوصافيحالةجمعويكيبيديا؟
3. ماهيأنواع الروابط التي يمكناستخدامهالتحسينأفضللنوعية نتائجالبحث؟
4. ما هي كيفيةحسابدرجةالروابط لمختلفالعناصرالمستخرجة كنتائج بحث؟هل ينبغيلناأننعتبر الروابط من نوع "ملف–ملف" أوبتعبيرأدقالروابط "عنصر–عنصر"؟ماهووزنالروابطالملاحيةمقارنةبالروابطالهيكلية؟
5. ماهوأثراستخداماالروابطفيالسياقالشاملأوالمحلي؟
6. ماهي كيفيةإدماجدرجةالروابطفياحتسابالنتيجةالنهائيةلعناصر XMLالمسترجعة في البحث؟
7. ماهوتأثيرنوعيةنتائج البحثالأوليةعلىسلوكالصيغالمقترحة؟

مقترحاتنا تحاول أساسا لمعالجة هذه المسائل:

أجرينا دراسة إحصائية عن نتائج البحث المسترجعة باستخدام نظام البحث عن المعلومات لجامعة داليانو التي أظهرت بوضوح أن الروابط تمثل علامة (إشارة) على أهمية العناصر في سياقالبحث عن المعلومات XMLباستخدام مجموعة الاختبار المقدمة من طرف INEX.أيضا، قمنا ببرمجة ثلاث خوارزميات تحليل الروابطو التي سمحت لنا بإجراء دراسة مقارنة تبين أن الطرق المعتمدة على الاستعلام (المحلية أو "الحساسة للموضوع") هي الأفضل مقارنة مع طرق السياق الشامل. اقترحنا خلال هذه الأطروحة ثلاثة صيغ لحساب نتيجة الروابط: الأولى تسمى " Topical Pagerank"، الثانيةتسمى "distance–based"، أما الثالثةفتسمى "weighted links based ". اقترحنا ثلاث صيغ دمج و هي : الصيغة الخطية، صيغة ديمبستر–شيفر والصيغة الغامضة.وأخيرا، أجرينا مجموعة من التجارب بما في ذلك معايير مختلفة. وقد أظهرت كل هذه التجارب أن: الطرق المقترحة قد حسنت دقة نتائجالبحث و ذلك في مختلف الصيغ المقترحة؛ الطرق "الحساسة للموضوع" هي الأفضل مقارنة مع الطرق ذات السياق الشامل. الطرق التي تستغل الروابط من نوع "عنصر– عنصر" تمكنت من الحصول على نتائج جيدة. صيغ الدمج التي تأخذ بعين الاعتبار الجوانب غير المؤكدة للنتائجالمحسوبة لعناصر XMLتسمح بتحقيق أداء جيد.

كلمات مفتاحية:

البحث عن المعلومات، XML، الطريقة القائمة على المسافة، الطريقة القائمة على وزن الروابط، تحليل الروابط، INEX، الدمج الغامض، الدمج ديمبستر–شيفر.

Dedication

To My Mother : Kheira

To My Wife : Khadidja

To My Dautghers: Asma & Meyssa

Acknowledgments

I would like to thank all those who helped me to make this thesis possible.

First of all, I express my cordial recognitions and gratitude to my two thesis advisors: Professor Mohamed MEZGHICHE and Professor Mohand BOUGHANEM, for their guidance and advice in the passing several years. I gratefully acknowledge them for time and attention they agreed to devote to the good progress of this work.

I am sincerely thankfulto the jury members who honour me by accepting to judge this modest work.

I want to thank all IRIT members, especially SIG team members, Karen Sauvagnat, Lynda Tamine, Ismail Badache, for their collaboration in my thesis project.

I would also like to thank all members of LIMOSE laboratory at UMBB University, particularly Dr. Rabah Imache.

I want to thank all Post-graduate service of UMBB University, especially Zineb Messaad for all the facilities in administrative procedures.

I am thankful to all my friends who encouraged and supported me throughout this work, in particular:Abdelghani, Sofiane, Faouzi, Farid, Tayeb, Lhouari, Reda and many others.

Finally, I have to thank all those whichin all good and bad times agreed to encourage me so that this work can be achieved.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AQE	Automatic Query Expansion
DB	DataBase
C S	Content And Structure
CLEF	Cross Language Evaluation Form
CML	Chemical Markup Language
CO	Content Only
CO+S	Content Only + Structure
DARPA	Defense Advanced Research ProjectAgency
DOM	Document Object Model
DTD	Document Type Definition
FLWR	For, Let, Where, Return
HTML	HypertText Markup Language
IDF	Inverse Document Frequency
IEF	Inverse Element Frequency
INEX	Initiative for the Evaluation of XML Retrieval
IQE	Interactive Query Expansion
IR	Information Retrieval
ITDF	Inverse Tag and Document Frequency
MATHML	Mathematical Markup Language
NEXI	Narrowed Extended XPath I
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OFX	Open Financial eXchange
OQL	Object Query language
PRP	Probability Ranking Principle
RF	Relevance Feedback
RSV	Retrieval Status Value
DBMS	DataBase Management System
SGML	Standard Generalized Markup Language
SMART	Salton's Magical Automatic Retriever of Text
SMIL	Synchronized Multimedia Integration Language
IRS	Information Retrieval System
TF	Term Frequency
TREC	Text Retrieval Conference
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XCG	eXtended Cumulative Gain
XFIRM	XML Flexible Information Retrieval Model
XML	eXtensible Markup Language
XML-QL	XML Query Language
XQL	XML Query Language

Introduction

The man, thirsty for knowledge, continues to conquer the fields of knowledge. In the days of *Denis Diderot*[1], the encyclopaedia was the holy work bringing the extract of human knowledge, but the century of light, in which we wrote about knowledge on soft paper, turned into an era of digitalization of all forms of knowledge. The areas of science, in perpetual expansion, have produced an informational flow unmanageable with conventional means, and now it seems advisable to use techniques to a lucrative management of human knowledge. Nowadays, and complying with the requirements of the technology, the digital documents are the main vehicle of information.

Indeed, the number of emails sent each day is around 165.8 billion (2014 statistics), and the total quantity of information produced each year would be around 912.5 exabytes[2], i.e. every second, 29000 Gigabytes of information are published worldwide.

"*More needles, more hay*", indeed, the more the amount of information increases, it becomes very difficult to satisfy a particular need. Information Retrieval (IR) is assigned the task of finding precisely the *needle in the haystack*. It is the implementation of automated tools for effective access to this huge amount of digital information. Information retrieval is concerned with the representation, organization, analysis, storage, access and presentation of information items (Lalmas, 2009). The information retrieval process is the process that allows to link all the information available on the one hand and the user's needs on the other hand.

I.1. Context

Our work is in the context of information retrieval, particularly information retrieval in semi-structured documents such as XML. XML (*eXtensible Markup Language*) is a standard format and universal data exchange known as XML which is a recommendation of the W3C (World Wide Web Consortium, February 1998) (Consortium, 1998). This standard format allows combining structural information named tags with content. This mixture between content and structure has led to the emergence of new IR challenges, particularly, in terms of querying.

Information retrieval in XML documents allows a user to express his needs by specifying the content he is looking for and/or the tag where he wants to find his response. Different types of queries can be built by combining or not keywords with tags, we often name these queries: Content Only (CO) queries, and Content and Structure (CAS) queries. Indeed, the hierarchical structure of XML documents is exploited as new evidence to retrieve XML elements at various levels of granularity. Therefore, and contrary to classical IR approaches that focus on searching unstructured content, (semi)structured IR combines both textual and structural information to answer queries. Several approaches exploiting the two types of evidences (textual and structural)

---

[1] (5 October 1713 – 31 July 1784) was a French philosopher, art critic, and writer. He was a prominent figure during the Enlightenment and is best known for serving as co-founder, chief editor, and contributor to the *Encyclopédie* (From Wikipedia).

[2] Statistics from : http://www.planetoscope.com/ (accessed in 2015)

have been proposed, they are mainly based on classic IR models that have been adapted to XML documents (Fuhr & Großjohann, 2001; Guo, Shao, Botev, & Shanmugasundaram, 2003; Kimelfeld, Kovacs, Sagiv, & Yahav, 2007; Mass & Mandelbrod, 2003; Theobald & Weikum, 2002).The XML structure has been used to provide a focused access to documents, by returning document component (e.g. sections, paragraphs, etc.), instead of the whole document in response to a user query.

In addition, the hierarchical structure of XML documents, they also may contain links connecting documents. These links have been widely exploited as an important source of evidence to search relevant pages in the context of Web retrieval.For instance, several algorithms were proposed, including PageRank (Brin & Page, 1998), HITS (Kleinberg, 1999) and SALSA (Lempel & Moran, 2001).

Despite their popularity in Web context, only few approaches, exploiting this source of evidence, have been proposed in the context of XML retrieval. These approaches can be classified into three categories: First, approaches which analyse the structure and nature of links in collections of XML documents. Second, approaches proposing link detection strategies, mostly under the "Link-The-Wiki" task of INEX campaign. These approaches focused on automatically linking orphan pages to already existing Wikipedia pages. Third, approaches exploiting links to re-rank the list of elements initially returned by an information retrieval system.In the context of this thesis, we focused on the third category of approaches.

The aim of our work is to propose approaches that exploit links as a source of evidence in the context of XML information retrieval.

I.2. Research Questions

The research questions addressed in this thesisare the following:

1. Can we consider the links as a source of evidence in the context of XML IR?

   According to many studies, link analysis seems to yield poor results in TREC's Web Ad Hoc retrieval task(Gevrey & Rüger, 2001). It is recommended to conduct a statistical study to provide some signs on the relationship between the links and the relevance of the elements.

2. Does the use of some well-known link analysis algorithms in the XML Information Retrieval context improves the quality of results, particularly in the case of the Wikipedia collection?

   The aim of studying this issue is to perform a comparative analysis between these algorithms and to seek any adaptation of these algorithms to the context of XML IR.

3. What types of links can be used to improve the relevance of retrieval results?

   Most of the proposed approaches in the literature only consider the navigational links, the hierarchical links (document structure) are generally ignored, while these links carries information that can be used effectively in the context of XML information retrieval.
   Moreover, in the literature there is almost no approach (excepting (Verbyst & Mulhem, 2009)) which allows taking into account the case of "*element-element*" links.

4. How to compute the link score of different elements returned as retrieval results? Should we consider the "document-document" or more precisely "element-element" links? What is the weight of the navigational links compared to hierarchical links?

   All these questions are related to the link score computation formula.

5. What is the impact of using links in the global or local context?

   Link analysis algorithms can be applied to one of two situations: global or local context. The global context (query-independent) in which the entire link graph of the collection is considered. Whereas, in the local context only a subset of the retrieved results are used to construct the link graph (query-dependent).

6. How to combine the link score and the content score of the returned XML elements?

   At this stage, a combination formula of the initial score with the link score should be proposed.Different parameters must be studied.

7. What is the impact of the initial results quality on the behaviour of the proposed formulas?

   The proposed formulas must be experimented on different retrieval systems according to their performance (top ranked system, mid-level system, etc.).

I.3. Contributions

Our proposals attempt essentially to address the various research questions expressed in the previous section:

- To answer the first question, in this case, *the possibility of considering the links as a relevance indicator*, we conducted a statistical study on the retrieval results returned by DALIAN IR system, using the IR test collection provided by INEX 2007. The study showed clearly that links are effective source of evidence and represent a sign of relevance of the elements in the context of XML IR.

- To answer the second question, *the capacity of somewell-known Web link algorithms to improve the retrieval accuracy in the context of XML Information Retrieval, particularly in the case of the Wikipedia collection*, we implemented three algorithms, i.e. PageRank, HITS and SALSA, in the *global context* (Pagerank) and *local context* (Topical Pagerank, HITS and SALSA). These implementations allow us to have a comparative study showing that *query-dependent* approaches (*topic-sensitive*) obtained the best performance. Our contribution in this stage is the "*Topical Pagerank*" implementation which is an adaptation of Pagerank to the local context (*topic-sensitive or query-dependent*).

- To answer the questions: 3,4 and 5, *type of used links (navigational and the hierarchical links), "document-document" or "element-element" links, global or local context,* we proposed three approaches:
  - The first is anadaptation of PageRank to the local context, called "*Topical Pagerank*" approach;
  - The second is the "*distance-based"* approach;
  - The third is the "*weighted links based"* approach.

These three approaches are "*query-dependent*", while the last two take into account the two types of "*element-element*" links:*navigational and the hierarchical.*

- To answer question 6, *the way the computed link score (or rank) is combined with the initial score to obtain the final score of the returned XML elements*, we proposed three combination formulas, i.e. linear formula, Dempster-Shafer formula and fuzzy-based formula. Different parameters have been studied at this stage (α parameter, discount rate, link score or rank, etc.).

- To answer the last question, *the impact of the initial results quality on the behaviour of the proposed formulas*, we conducted a set of experiments consisting of:
    - Firstly, using of retrieval results from different XML IR systems (top-ranked, mid-ranked systems)
    - Secondly, using retrieval results related to different collection. In our case, we have experimented based on the retrieval results of the INEX 2007 and INEX 2009 test collections.

  All these experiments have shown that:
    - the proposed approaches have improved the retrieval accuracy for the different tested configurations (for top-ranked and also for mid-ranked systems);
    - the "*topic-sensitive*" approaches are the best compared to global context approaches;
    - approaches exploiting the links of type "*element-element*" obtained good results;
    - The combination formulas taking into account the uncertain aspects of computed scores of XML elements allow achieving good performance.

I.4. Publications as part of the thesis

We published as part of this thesis the following list of journal and conference papers:

1) M'hamed MATAOUI & MEZGHICHE Mohamed, « *Prise en compte des liens pour améliorer la recherche d'information structurée* », Actes de la sixième conférence francophone en recherche d'information et applications (**CORIA 2009**), Presqu'île de Giens, France, Mai 2009. pp. 363-372.
2) M'hamed MATAOUI, Mohand BOUGHANEM & Mohamed MEZGHICHE, « *Experiments on PageRank Algorithm in the XML Information Retrieval Context* », In Proceeding of the Second International Conference on the Applications of Digital Information and Web Technologies (**ICADIWT 2009**), London, UK, August 2009. pp. 393- 398. ISBN: 978-1-4244-4457-1

3) M'hamed MATAOUI, Mohamed MEZGHICHE & Mohand BOUGHANEM, « *Exploiting Link Evidence to Improve XML Information Retrieval* », Actes de la première édition de la conférence maghrébine sur l'Extraction et la Gestion des Connaissances (**EGC-M 2010**), Alger, ALGERIE, Décembre 2010. pp. 23-33.

4) M'hamed MATAOUI, Faouzi SEBBAK, Farid BENHAMMADI and Mohamed MEZGHICHE, « *Evidential-Link-based Approach for Re-ranking XML Retrieval Results* ». In Proceeding of the **DATA 2014** Conference, Vienna, Austria, 2014. pp. 64-71. ISBN: 978-989-758-035-2

5) M'hamed MATAOUI, Faouzi SEBBAK, Farid BENHAMMADI and Kadda Beghdad Bey. «*A Fuzzy Link-Based Approach for XML Information Retrieval* ». In Proceeding of the IEEE International Conference on Fuzzy Systems (**FUZZ IEEE 2015**), Istanbul, Turkey, 2015.

6) M'hamed MATAOUI and Mohamed MEZGHICHE. « *A distance based approach for link analysis in xml information retrieval* ». In: **Computer Systems Science and Engineering**. Volume 30, Issue 3, pp. 173-183. CRL PUBLISHING LTD, 2015. (***ISI Thomson***)

I.5. Organization of the thesis

This thesis is organized into sixchapters:

Chapter I,***Basic ConceptsofInformation Retrieval***,describes general concepts ofInformation Retrieval. The first sectionis devoted to thedescription of the IRprocess(SectionI.2), in which we definethe notions ofcollection of documents (Section I.2.1), information need (Section I.2.2), indexing(SectionI.2.3), query-document matching(sectionI.2.4)andquery reformulation(SectionI.2.5). In sectionI.3, we surveyIR models. Finally, SectionI.4discusses the evaluationmethodology inIR.

Chapter II,***XML Information Retrieval***, overviews presentsa state of theartofXMLInformation Retrieval (XML IR). SectionII.2shows thesemi-structured/XML documents while definingthe notion ofstructure.SectionII.3lists thespecific problems ofXMLIR relative to information granularity, indexing strategies and query processing. SectionII.4presentsthe different techniquesfor indexingXMLdocuments.SectionII.5 provides an overview ofquery languagesthat take into accountthe structural aspectofXML documents.SectionII.6presents a surveyof the XML retrieval models. SectionII.7shows three examples ofXML IR systems. Finally, sectionII.8is devoted to the evaluationof XML IR.

Chapter III, ***Links in Web Information Retrieval***, we will review the main concepts, and approaches proposed in the Web search, in particular the use of hyperlinks as a source of evidence in the retrieval process. SectionIII.2 presents the concepts of Hypertext and the Web information retrieval field. SectionIII.3 describes the value of link evidence in the IR field. SectionIII.4 describes an outline of some well-known link-based ranking algorithms (mainly used in the Web IR), namely, PageRank, HITS and SALSA. SectionIII.5 presents some aspects related to links and search engines. SectionIII.6 will conclude this chapter with a brief discussion.

Chapter IV, ***State of the art on the use of links in XML IR***, gives an overview of the use of link evidence in XML IR context. Section IV.2. is devoted to the description of the approaches studying the structure and the nature of INEX Wikipedia links. In sectionIV.3 we review some link detection approaches. SectionIV.4discusses approaches using the re-ranking principal based on link evidence.

Chapter V, ***Proposed Approaches***, detailsour approaches for the exploitation of link evidence in the context of the XML information retrieval. We start this chapter by a statistical study (sectionV.2) where we evaluate if links are an effective indicator of relevance in the case of Wikipedia corpus. Next, in the second part of this chapter (sections:V.3, V.4 and V.5), we focus on the way we compute the link score of retrieved XML elements by detailing our three formulas of link Score computation. In the third part of this chapter, we describe the three approaches of combining link score and content score we proposed (sectionV.6).

Chapter VI, ***Experiments, Results and Discussion***, presents the results of the experiments we conducted to evaluate the different proposed approaches. SectionVI.2, describes the experimental setup and evaluation protocol. The experimental setup will include the details of test collections, tools and evaluation measures. The evaluation protocol (sectionVI.3) define the way the experimental results have been obtained. SectionVI.4, describes and discusses the results obtained by our different proposals.

We finish our thesis with a general conclusion thatsummarizes the main contributions of this work and provides some future works.

Chapter I.

Basic Concepts of Information Retrieval

I.1. Introduction

Information Retrieval (IR) is a field of the data processing which deals with the representation, organization, analysis, storage, access and presentation of information that will satisfiesthe information need of a user(Baeza-Yates & Ribeiro-Neto, 1999; Salton, 1971; Salton & McGill, 1986). The information need is usually expressed in natural language, not always well structured and often semantically ambiguous. The IR field handles various concepts, such as query, information need, documents, and relevance.

Several definitions were given in computer science literature to the concept of IR:

"*Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information. It is another, more general, name for the production of a demand bibliography. Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation. Information retrieval is crucial to documentation and organization of knowledge*"(Mooers, 1951).

"*Information retrieval,(IR) part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language.*"(Baeza-Yates & Ribeiro-Neto, 1999).

This chapter focuses some important concepts ofIR. It is organised as follows: the first section is devoted to the description of the IR process (sectionI.2), in which we define the notions of collection of documents (section I.2.1), information need (section I.2.2), indexing (section I.2.3), matching functions (section I.2.4) and query reformulation (section I.2.5). In sectionI.3we review some IR models. SectionI.4discusses the evaluation process in IR.

I.2. Information Retrieval Process

Given a collection of documents, information retrieval process is designed with the objective of providingthe list of documents that would contain the information that answers the user information need expressed by a query. Most of IR systems follow the same processing steps. Documents are first indexed to build the list of keywords that represent a document, to answer a query. This latter is first proceeding to extract keywords, than it is compared to the stored index. The concept of relevance is strongly subjective, because it depends on the user, therefore very difficult to automate.The IR process includes several concepts:

- collection of documents;

- The information need;
- The indexing function;
- The "query-document" matching function;
- The relevance feedback function.

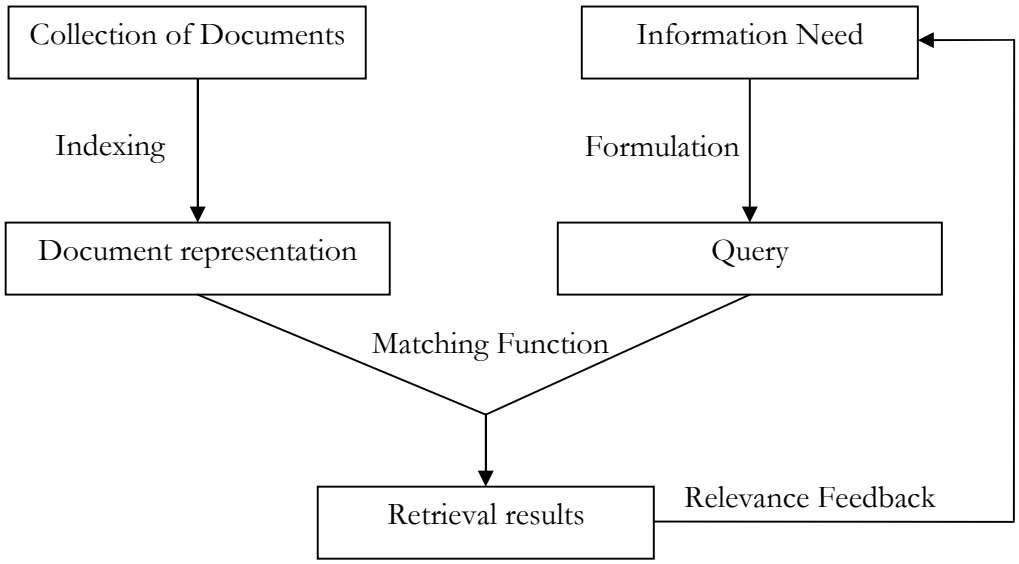


Figure I.1:Information Retrieval Process(Rijsbergen, 1979)

I.2.1. Collection of documents (corpus)

A corpus of documents is a set of documentary unit which can be whole documents or many parts of documents. In the classic IR the unit of information used and sought during the retrieval process is the document.

I.2.2. Information Need

Information need is the mental expression of what the user seeks. The expression of a need is done by a query which allows the interrogation of IR system. The interrogation of anIR system can take several forms, namely: interrogation in natural language, interrogation in Boolean language or graphic interrogation. According to Kleinberg(Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999) there exists three various forms of queries: a) specific queries; b) largequeries; c) queries by similarity.

Generally, queries are made up of a set of keywords. These keywords can be connected to each other by Boolean operators, as they can be also organised in the form of expressions.

Another type of queries is that specific to the structured/XMLIR which takes into account structural constraints in addition to the textual information. This type of queries will be detailed in chapter 2.

I.2.3. Indexing Function

Indexing is the process of creating a representation of a document and queries that can be easily and efficiently manipulated by anIR system in terms of space (document storage) and time (retrieval process) requirements. This representation, called "Index", is a data structure built from the text to speed up the searches (Baeza-Yates & Ribeiro-Neto, 1999). Indexing consists in analysing the documents to extract a set of keywords used as descriptors of these documents. There are three types of indexing:

a) Manual Indexing.

The extraction and the choice of descriptors are done by a librarian or a specialist.

b) Automatic Indexing.

The extraction and selection of descriptors take place in a fully automated manner.

c) Semi-Automatic Indexing.

The extraction of descriptors is performed by the system and the choice of descriptors is left to the specialist.

A comparative study of automatic and manual indexing was done by Anderson and Perez-Carballo (Anderson & Pérez-Carballo, 2001). The results show that the advantages and disadvantages of both indexing methods tend to balance out. In other words, the choice of one or the other depending on the field of interest, the collection and the application considered.In the following sections we will describe in detail the various steps of the automatic indexing from extracting document keywords to the index construction.

The efficiency of an index based IR system can be measured by many factors: a) indexing time; b) indexing space; c) index storage; d) query latency; e) query throughput;

I.2.3.2. Lexical analysis/Tokenization

The lexical analysisphase is used to extractall the termsbelonging to adocument.This is carried outby taking into account theseparation spacesbetweenwords, numbersandpunctuation.A term canbea single keyword(eg. apple)ora compound keyword("information retrieval") butsimple wordsare often used in IR.

I.2.3.3. Stop-wordsRemoval

Stop-words arewords whichhave no value for retrieval purposes andcarrylittle meaning, they increase the size of the indexandthey make the searchslower.Therefore, their removal is oftenimperative. Stop-words removal is importantinsofar asitis a factorwhich has a greatinfluence on the retrieval accuracy.Indeed, thefailure toremovethe stop wordsinevitablycauses retrieval noise(irrelevant retrieved documents).If a queryincludes the word "of" thenall documents in thecorpusmatch! There are twomost known techniques for removingstop words: a) the use of astoplist containing: articles, prepositions, conjunctions, pronouns…; b) removal of words havingtheir occurrences in the collection greater than a defined threshold.

I.2.3.4. Stemming

Stemming is the process of reducing inflected forms of words to their grammatical root. It aims at replacing all the variants of a word with the single stem of the word. A given word can have several forms in a text whose meaning is almost similar. Several stemming methods have been proposed in the literature, including : the truncation method (Frakes, 1992), n-gram method (Adamson & Boreham, 1974), dictionaries or affixes elimination methods (e.g. Porter algorithm (Porter, 1980).

Lemmatisation thus increases the recall, which reduces in practice the accuracy rate. This is due to the loss of the original semantics of the word during the transition to the final form.

I.2.3.5. Term Weighting

Term weighting phase is used to measure the importance of a term in a document. The importance of a term is usually measured by statistical or linguistic methods.

Most weighting schemes proposed in the IR literature are based on two factors, namely: local weighting and global weighting(Robertson & Jones, 1976; Salton, 1971). The first quantifies the local representation of a term in the document (*tf: Term Frequency*), and the second quantifies the representationin thecollection of documents (*idf: Inverse of Document Frequency*).

- ***tf (Term Frequency)***: This measure indicates the importance of the term in the document. This importance is proportional to the frequency of the term. Several local weighting formulas have been proposed, including: number of occurrences function, the binary function, logarithmic function and normalized function(Robertson & Jones, 1976).

- ***Idf (Inverse of Document Frequency)***:This factor measures the importance of a term in the whole collection (global weighting). It reflects the impact of a term according to its number of occurrences in the collection. The formula that expresses the importance of a term in the collection can be seen as follows: *log (N/df)*, where *df* is the number of documents containing the term and *N* is the total number of documents in the collection.

The combination of the two measures (*tf*and *idf*) gives a good approximation of term importance in a document, particularly in homogeneous corpus ofdocuments. The weighting functions are often referred to as *tf-idf*schema.

Another factor has been proposed to mitigate the negative effects of the size of the documents on the *tf* factor. Robertson(Robertson & Walker, 1994) and Singhal et al. (Singhal, Salton, Mitra, & Buckley, 1996)proposed to incorporate the size of the documents in the weighting formula , this factor is called normalization factor (see formula).

$$finalweight(t) = \frac{tf \times idf \quad weight}{euclideanlength\,of\ \ documentvector} \tag{I.1}$$

I.2.3.6. Index Construction

The index, defined as the storage structure used to store informationselected during indexing process.This structure is used to select for any term all documents where this term appears. Several storage solutions have been proposed including:*inverted files*, *signature files* and t*suffix arrays*.

I.2.4. Query-Document Matching Function

This function is definedto measure therelevance of a documentwith respect to a query.It is seen asa probabilityor vectorsimilarity, denoted *RSV(Q,d)*(Retrieval StatusValue), where*Q*represents the queryand*d*represents a document. Thesimilarity functionallows rankingthe retrieved documents by relevance.

I.2.5. Query Reformulation

Building a query that represents accurately the user information need is often a difficult task. Therefore, the documents retrieved by the initial query may not fulfil the information need of the user. Query reformulation aims at building a new query that better fits user information need. Query reformulation can be performed by two strategies: query expansion by adding new terms, and/or reweighting the initial query terms. Query modification can be manual, automatic or semi-automatic.

I.3. IR Models

The role ofan IR model is to provide a formalization of the information retrieval process. The most important role is to provide a theoretical framework for modelling the relevance measure(Salton, Fox, & Wu, 1983). There are three mainclasses of models according to their mathematical basis (Kuropka, 2004):

- Set-theoretic models;

- Algebraic models;

- Probabilistic models.

The following figure presents the taxonomy of IR models according to (Kuropka, 2004).

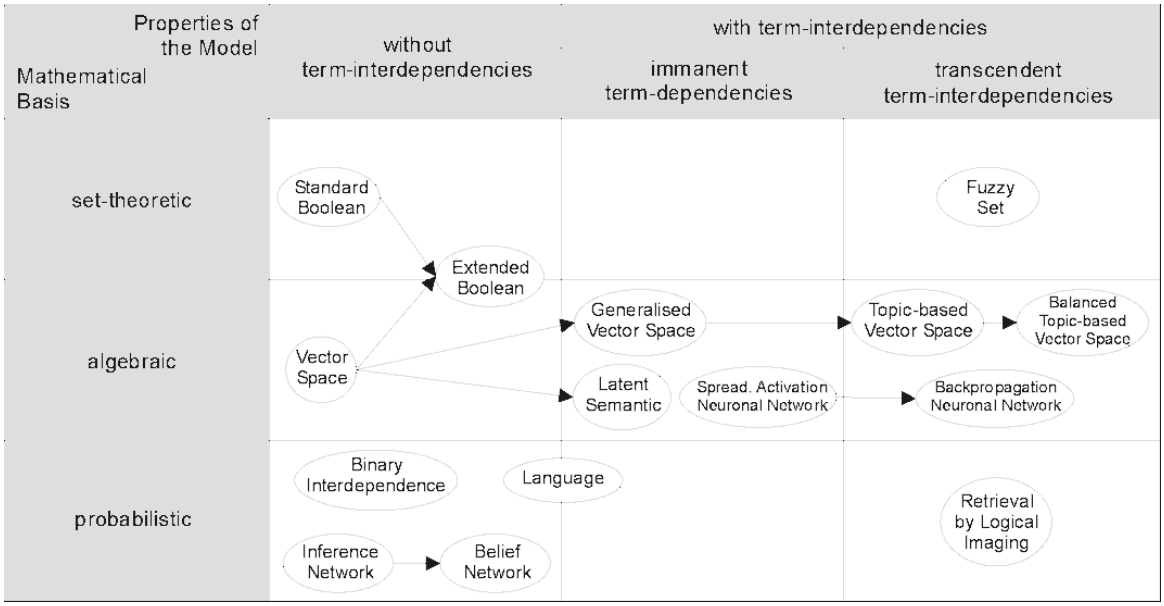


Figure I.2 : ClassificationofIRmodels (Kuropka, 2004)

Theset- theoretic models are based on set theory. In these models,thequery termsare separatedby logical operators: conjunction (AND), disjunction (OR) andnegation(NOT). These operatorsare used to performunion"OR" Intersection"AND"and difference"NO"operationsbetween theresult setsassociated with eachquery term.

Algebraic models are based on the algebraic theory. In these models, the documents and queries are represented as vectors, the relevance of a document with respect to a query is defined by distance (or similarity) between their associated vectors.

Finally, probabilistic models are based onprobability theory. For these models, therelevance of a document with respect to a query isseen asa probability ofrelevance ofthe document (resp. query) with respect to a query (resp. document).

In the following sections, we describe for eachof these classes the most representativemodel(i.e.theBoolean model, the vector space modelandthe probabilistic model).

I.3.1. Boolean Model

The Booleanmodel isthe first model thathas emergedin the IR field.It is based onthe setstheory andBoolean algebra. In this model,a queryis represented by a logical expressioncomposedof terms separated bylogical operators(AND, OR and NOT).Term frequencies in the index, i.e., term-document matrix, are all binary,that is tosaythat the terms arepresent or absentin the document ($w_{ij} \in \{0,1\}$). The Booleanmodel usestheexact matching method, i.e., only documents with respect to the describedquery.The similaritybetween a document andaquery is definedby:

$$\begin{cases} RSV(q,d) = 1 \quad if\ d\ belongs\ to\ the\ set\ described\ by\ the\ query \\ \qquad\quad 0 \quad else \end{cases} \tag{I.2}$$

Thus,a document is consideredasrelevantorirrelevant.The results ofthe similarity functiondonot allowto returna rankedlist of documents.

I.3.2. Vector Space Model

The vector space model is one of the first models based on the statistical approach. It consists at representing documents and queries as vectors of weighted terms. The basic idea of that model is to use a geometric representation to rank documents by their relevance with respect to a query, i.e., documents and queries are represented as vectors in vector space generated by the extracted terms of all documents in the collection. This idea was developed by Gerard Salton and his team (Salton, 1971) in their SMART (*Salton's Magical Automatic Retriever of Text*) project. Salton proposed a model based on a similarity measure by the scalar product.

Unlike theBoolean modelwherethe query termsmust be connectedby logicalconnectors,the vector space modelallows the userto express hisinformation needin the form ofa listof keywords ornatural language.

Formally,thevector space model, therepresentation of a documentis seen asa vector $\vec{d}_j = \{w_{1,j}, w_{2,j}, \ldots, w_{t,j}\}$where $w_{i,j}$is the weightof the termsin the document,*t* isthe total number

ofindex terms, and the queryisalsoseen asa vector $\vec{q} = \{w_{1,q}, w_{2,q}, \ldots, w_{t,q}\}$.One of the simplestmeasuresof similarity isthe scalar value:

$$RSV(\vec{d}_j, \vec{q}) = \sum_{i=1}^{t} w_{i,j} * w_{i,q} \tag{I.3}$$

Thissimilarity measureis to measurethe number of sharedtermsbetweenthe queryand documents, as the weightof the termsare binary.

Severalsimilarity functionshave been proposedin the literature.Weincludethe most commonfunctions:Cosine, JaccardandDice measures.Cosine measure is the most used in IR.

Cosinemeasure: $$Sim(D_j, Q_k) = \frac{\sum_{i=1}^{N} (wd_{ij} * wq_{ik})}{\sqrt{\sum_{i=1}^{N} wd_{ij}^2 * \sum_{i=1}^{N} wq_{ik}^2}} \tag{I.4}$$

The advantagesof that vector space modelare many:it allowsterm weighting, which increases system performanceand allowsreturning documentsthat best match(matchapproximately)queryin decreasing order of their relevance with respect to a query.Thedocuments can bereturnedin descendingorder of theirdegree of similarity to the query.Greater the degree ofsimilarityof a document, the higher the documentcorrespondsto the queryanditis likely to berelevant to the user.

Theoretically, the vector model has the main disadvantage of the mutual independence of indexing terms. Wong et al. (Wong, Ziarko, & Wong, 1985) proposed a "Generalized Vector Space Model", which removes the assumption of word independence.

I.3.3. Probabilistic Model

The firstprobabilistic modelwas proposed byMaronandKuhns (Maron & Kuhns, 1960)in the early1960s. The basic principleof the probabilistic modelis to present theIR system retrievalranked based onthe probability ofrelevance of a documentwith respect to a query.Robertson(Robertson, 1977)summarizes thecriterionof orderbythe "principle ofprobabilisticclassification",also referred to asPRP (Probability RankingPrinciple).

In the probabilistic model, answer a query is equivalent to specifying the properties of a set of documents called "ideal answer set". This set containsexactlythe relevant documentsand no others.At the time query, the properties of the"ideal answer set"are not known;there must be afirst attemptto generatea firstprobabilistic descriptionofthis set.Then,interact with the user must take place to improve the probabilistic description(Robertson, 1977).

To measuretherelevance,the probabilistic modelis based on thedistributionof terms ina representative sample oflearning set documents. Theassumptionsare:

- termsdistributionin the relevant documentsis the same astheirdistributionoverall documents;
- "relevantdocument","documentirrelevant" are independent variables.

The retrieval processis reflected in thecomputation of the relevancedegree(orprobability) ofdocument with respect toa query.Two conditionalprobabilities are usedinthe decision process: $P(w_{ij}/Pert)$: probability that termt_i occurs in thedocument d_j knowing that$d_j$is relevant tothe query. $P(w_{ij}/NonPert)$: probability that term$t_i$occurs in thedocument d_j knowing that$d_j$is not relevant tothe query.

The probabilistic model has been implemented by Robertson and Walker (Robertson & Walker, 1994; Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1995) in the Okapi system.

I.3.4. Other Models

I.3.4.1. Language Models

Thelanguage modelsare probabilisticmodels.Theydenotea probability functionthat assigns probability to eachword sequence. Their objective is tocapture linguisticregularitiesof a language by observing the distribution of words, sequences ofwordsin a givenlanguage.

TheIR based language modelsare based onthe following hypothesis:"a user interacting with an information retrieval systemprovidesa query thinking at one or more documentshe desires". In a languagemodel a document is consideredrelevant whentheuser queryis similar to thatinferred fromthe document.The principleis toestimate the probabilitythat a queryisinferred fromthedocument (Boughanem, Kraaij, & Nie, 2004). This probability(denoted$P(Q_k|D)$) will be used to rank thelist ofretrieved documents. The following measureallows the ranking of retrieved documents:

$$P(T_1,T_2,...,T_n \mid \mathrm{D}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(\mathrm{T_i} \mid \mathrm{D})) \tag{I.5}$$

Such as T_1, T_2, …, $T_n$are randomly independent variables representing thequery terms, andλ_i estimates the importance ofthe querytermTi. $P(T_i|D)$represents the relevance probability ofthe termT_i in thedocument$D$:

$$P(Ti \mid D) = \frac{tf\,(Ti \mid D)}{\sum_T tf\,(T, D)}\ , \tag{I.6}$$

Where$tf(T_i|D)$is the $T_i$termfrequency inthe document$D$.

I.3.4.2. Bayesian Model

The use of Bayesian networks in the IR field was introduced by Turtle and Croft (Turtle & Croft, 1989). It belongs to the family of probabilistic models. Conceptually, Bayesian networks use directed graphs to show probabilistic dependencies between variablesand have led to the development of sophisticated algorithms for propagating influence so as to allow learning and inference with arbitrary knowledge within arbitrary directed acyclic graphs. Turtle and Croft used a sophisticated network to better model the complex dependencies between a document and a user's information need.

Regardless ofthe attractiveness of an IR model with firm theoretical foundations, existing parameters estimation methods are to a certain degree unsatisfactory. Since parameters estimation requires relevance judgments for the present query, another method of producing the initial document ranking must be used.

Bayesian IR model overcomes some of the weaknesses of existing probabilistic models (poor use of relevance feedback, judgements for past queries). It has the following strengths: first, it retains the thorough theoretical foundations of the probabilistic models with the capacity to produce an initial document ranking; second, it provides an automatic mechanism for learning; third, it allows incorporating relevance information from other queries.

The parameters of the Bayesian IR model are: $\pi_{R_i} = P_q(t_i|R)$and $\pi_{\bar{R}_i} = P_q(t_i|\bar{R})$, $i = 1,\ldots,p$. Rather than ad hoc techniques, prior distributions $p(\pi_{R_i})$ and $p(\pi_{\bar{R}_i})$, $i = 1,\ldots,p$, are assessed for these parameters. These prior distributions represent the prior knowledge of the query-document associations. This prior knowledge may be obtained from relevance data based on past queries.

The initial probabilities of relevance for each document are computed using the expected value of the prior distributions to estimate the model parameters. IR system present a first documents ranking to the user and ask for relevance assessments. These relevance assessments are used as query-specific data to learn by interaction with the user,$X_r = (X_{r1}, X_{r1}, \ldots, X_{rn})$, that will be used to modify the distribution on the model parameters to obtain: $p(\pi_{R_i}|X_r)$ and $p(\pi_{\bar{R}_i}|X_r)$. The following equation represents a formulation of the Bayesian IR model:

$$p\left(n_{R_i}|\pi_{R_i}\right) = \binom{r_k}{n_{R_i}} (\pi_{R_i})^{n_{R_i}} (1 - \pi_{R_i})^{r_k - n_{R_i}} \tag{I.7}$$

Where:

- n_{R_i}represents the number of occurences of term I in the set of relevant documents.
- $r_k$represents the number of documents that the user assessed as relevant.

The implementation of the Bayesian IR model must consider two main issues:

- First, the specification of prior distributions, by answering the question: how to incorporate the prior knowledge about the model parameters in the prior distributions?
- Second, updating the distributions, by answering the question: which, and how many, documents should we present to the user for relevance feedback?

I.3.4.3. Logic-Based Models

The time spent by indexing and retrieval processes is an important parameter in IR systems. Classical IR approaches are mainly guided by efficiency rather than expressiveness. The Boolean and the VSM models retrieve efficiently documents but the expressiveness of the representations of documents and queries is poor.

Few attempts aiming to increase the expressiveness of the representations of queries and documents have been made. Two main points regarding expressiveness must be taken into account when proposing an IR model: first, efficiency of the indexing and retrieval processes must be assured; second, methods that generates automatically the representations of the

documents must be specified.The semantics of logic provides the ability to write document's representations which capture the document content in a better way.

The basic point of the logic-based IR models is the assumption that documents and queries can be represented effectively by logical formulas (see Figure I.3). The basic logical test is to decide whether or not the formula representing a query can be inferred from the formula representing the document. Information items can be modelled with a rich and uniform framework provided by logic. Indeed, some logical-based models express with a homogeneous framework many IR features, such as: links, multimedia content and user's knowledge.

Documents and queries are represented by logic formulas and the notion of logical consequence is exploited to decide relevance, i.e. a document d is relevant to a query q iff $d \models q$, where $\models$ is the classical entailment, that consists on evaluating whether or not every model of d is a model of q.

Classical Representation	**Logical Representation**
Indexing Vocabulary: $\{t_1,t_2,t_3,t_4\}$	Propositional Alphabet: $\{t_1,t_2,t_3,t_4\}$
Document: (0,1,0,1)	$\neg t_1 \wedge t_2 \wedge \neg t_3 \wedge t_4$
Query: (1,0,0,1)	$t_1 \wedge t_4$

Figure I.3 : An example of translation from binary-weighted vectors to logical formulas

I.4. Evaluation of IR Systems

IR approachesattemptto bringtherelevance of the system (represented by the results retrieved by the IR system) as close as possible tothe user relevance (user satisfaction in these results).Although theresponse time andthe used space forthe indexing of informationare more or lessimportantin the evaluation ofIR systems, the retrieval accuracy(also calledeffectiveness)remains the mostimportant criterion.Effectiveness is often evaluated using two criteria: recall and precision. Thisevaluation phaseallowsusually comparing between IR systems.

One of the hardest tasks in evaluation is to identify the ground truth, the relevant documents of queries. To cope with this problem, most of the IR evaluation methodologies are based on Cranfield paradigm (Cleverdon, 1967). This paradigm suggests to conduct evaluation (compare IR approaches) by using a test collection composed of:

- Document set;
- Queries (topics);
- Relevance assessments (judgements).

Several evaluation campaigns have been created (TREC3, INEX4, CLEF5 …) (Gövert & Kazai, 2002; D. K. Harman, 1993; Peters, 2001).

We describe below the most common parts of IR systems evaluation.

I.4.1. Test Collections

Since the 70s, several projects aiming at creating test collections and evaluation protocols have emerged. These projects provide a framework for the comparison of different IR systems and algorithms. They have provided several reference collections in IR: the CACM collection, Cranfield collection, GOV2, NTCIR, the CLEF collection (*Cross Language Evaluation Form*) and the ISI collection.

One of the most important projects of IR evaluation was initiated by DARPA (*Defense Advanced Research Project Agency*) co-organized with the NIST (*National Institute of Standards and Technology*). This project is called: TREC Evaluation campaign (*Text REtrieval Conference*) (D. Harman, 1993). The campaign started in 1992; its purpose is to encourage the information retrieval field on large test collections. It provides a set of documents and queries to participants who will then forward the retrieval results of their IR systems to the NIST. NIST selects the *N* first documents of each IR system and provides them to the assessors who decide the relevance of each document.

Initially, TREC had two tracks: the ad hoc track and the routing track. Thereafter, new tracks have appeared, for instance, Web, video, terabyte, robust retrieval, relevance feedback, question answering, interactive, cross-language tracks. The tracks of TREC 2013 (NIST, 2013) are : contextual suggestion, crowdsourcing, federated Web search, knowledge base acceleration, microblog, session, temporal summarization, Web tracks.

I.4.2. Queries

An important component of IR tasks is the definition of information need. Queries represent the expression of the user information need in the input language provided by the IR system. The user specifies his information need through a query (set of connected terms) which initiates the retrieval process for extraction of the relevant documents. The most common type of input language allows simply the specification of keywords and of a few Boolean connectives (Baeza-Yates & Ribeiro-Neto).

I.4.3. Relevance Assessments

Relevance is an important concept in IR that has been widely debated and considered. It is subjective to assess, i.e., documents considered as irrelevant in the ad hoc retrieval task could possibly be relevant for other retrieval tasks (home page finding). Relevance could have different

[3] trec.nist.gov

[4] https://inex.mmci.uni-saarland.de/

[5] http://www.clef-initiative.eu/

interpretations. Relevance denotes how well a retrieved set of results (of all information forms) meets the information need of the user.

Cosijn and Ingwersen (Cosijn & Ingwersen, 2000) distinguish five expressions of relevance: algorithmic, topical, cognitive, situational and socio-cognitive. In (Saracevic, 1975),the author describes a framework for thinking about relevance. He distinguishes several different interpretationsof the relevance concept. He describes the relevance concept measure of the effectiveness of the contact between a source and a destination in a communication process. A large overview of several decades of research related to relevance can be found in Mizzaro (Mizzaro, 1998).

Some other interpretations of the relevance concept can be found in the IR literature:topical relevance, system relevance, user relevance, pertinence, utility, situational relevance.

I.4.4. Evaluation Measures

I.4.4.1. Recall andPrecision

The precision and recall measures allow evaluating the ability of an IR system to answer two main objectives: find all relevant documents and reject all non-relevant documents. To present these two measures, we introduce (see following figure) partitioning of all returned documents by the IR system (denoted B) in two subsets: a subset of relevant documents and a subset of irrelevant documents.

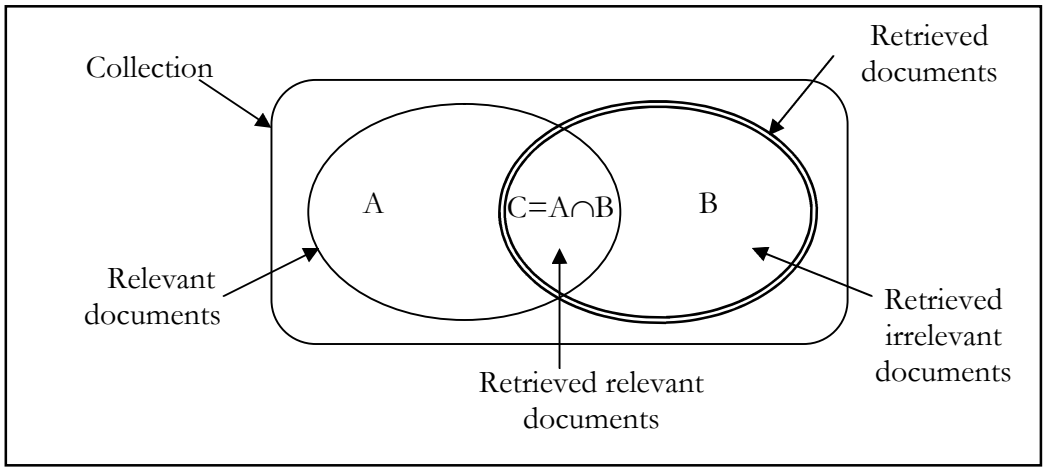


Figure I.4 : Partitioning of all documents for a query

Therecall rateand precisionare defined as follows:

Recall:measures the abilityof the systemto retrieveall relevant documentsin response to aquery.In other words,it measures theproportionof relevant documentsthat are retrieved. It is expressedby the following formula:

$$\text{Recall} = \frac{|C|}{|A|} \tag{I.8}$$

Precision:measures the abilityof the system torejectall irrelevant documents to a query.In other words,it measures theproportionof retrieved documents that are relevant.It is expressedby the following formula:

$$\text{Precision} = \frac{|C|}{|B|} \tag{I.9}$$

Precision-Recall Curve

To examine effectively the results, we must calculate the pairs of measurements (recall rate, precision rate) at each retrieved document. To illustrate the computation of recall and precision, we give an example (Table I.1) which represents the retrieval results returned for a query (*Q1*) by two systems (*S1, S2*) in a collection containing 10 relevant documents. The precision-recall curves are plotted associated in Figure I.5.

We noticethat for the same recall pointcorrespond many precision values.One manner to makeit easier toreadthe "precision-recall" curvesis torepresentthe computedprecision at each recall point.

We say that asystem is perfectif itreturns only the relevant documents, witha recall anda precision equal to 100%. In practice, the two measures vary inversely, the precisiondecreasesasthe recallincreases, whichmeans that therecall-precisioncurveis often decreasing.Asystem *A*is said toperform better thansystem *B*ifitsrecall-precisioncurveis on top comparedto that of*B*.From theFigure I.5we can deduce thatthesystem *S2*is better thanS1.

	S1 IR system			S2 IR system		
Rank	relevant/irrelevant	recall	precision	relevant/irrelevant	recall	precision
1	√	0.100	1.000	√	0.100	1.000
2	X	0.100	0.500	√	0.200	1.000
3	X	0.100	0.333	X	0.200	0.666
4	√	0.200	0.500	√	0.300	0.750
5	√	0.300	0.600	X	0.300	0.600
6	X	0.300	0.500	√	0.400	0.666
7	X	0.300	0.428	X	0.400	0.571
8	√	0.400	0.500	√	0.500	0.625
9	X	0.400	0.444	√	0.600	0.666
10	X	0.400	0.400	X	0.600	0.600

TableI.1 : Example of "Precision-Recall" computation for the two IR systems *S1* & *S2*

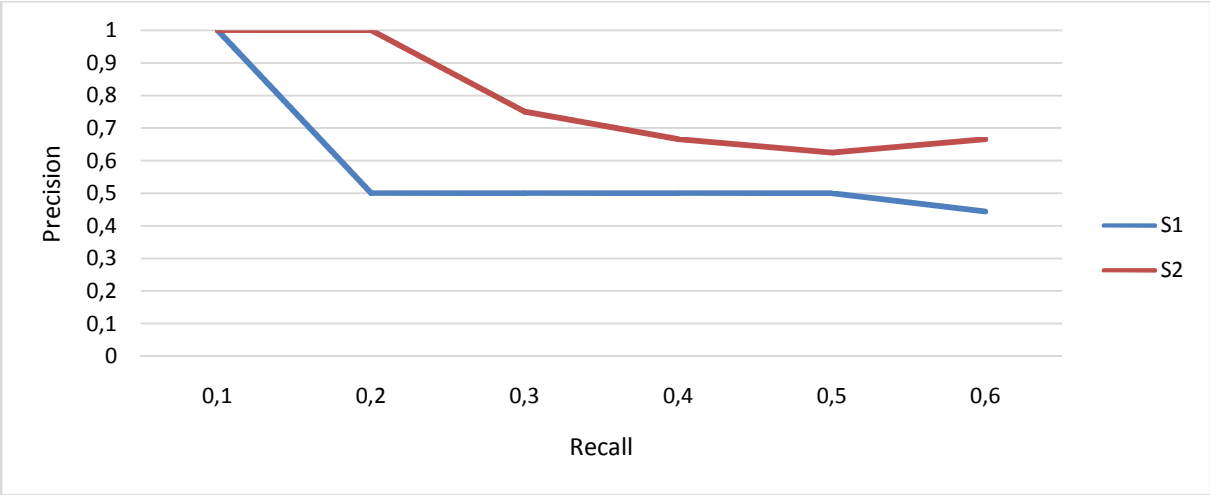


Figure I.5 : "Precision-Recall" Curves for the S1and S2retrieval systems

I.4.4.2. Other Evaluation Measures

Basedon the principleof therecall and precision measures, other measuresattempt tocombine recall and precisionin order toobtaina single value.Amongthe proposed measureswe may mention:F-measure, average precision, mean average precision and discounted cumulative gain.

a) Harmonic measure

The harmonic measure H is a function that combines the two values of recall and precision into a single value included between [0,1] (Shaw Jr, Burgin, & Howell, 1997).

$$H(j) = \frac{2}{\frac{1}{R(j)} + \frac{1}{P(j)}} \tag{I.10}$$

Where:

- *R(j)* and *P(j)* represent respectively recall and precision of the j^{th} retrieved document by the IR system.

This measure tends to 0 when no relevant documents are returned and equal to 1 when all the retrieved documents are relevant.

We can observe that the function *H* takes high values when the values of recall and precision are high.

b) « E » measure

This measure was proposed by Van Rijsbergen (Rijsbergen, 1979). Its purpose is to allow the user to specify which of the values of precision or recall is more interesting. The measure is defined by:

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{R(j)} + \frac{1}{P(j)}} \tag{I.11}$$

b is a parameter that allows the user to specify the importance of recall and precision. If *b* is equal to 1, *E(j)* will take the value of the complement of the harmonic measure *H(j)*. If (*b*<1), the recall is privileged and if (*b*> 1), the precision is privileged.

c) NDCG

NDCG (*normalized discounted cumulative gain*) is designed for situations of non-binary notions of relevance. Like precision at *k*, it is evaluated over some number k of top search results. For a set of queries *Q*. let *R(j, d)* be the relevance score assessors gave the document d for query *j*. Then,

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)} \tag{I.12}$$

Where Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's*NDCG* at k for a query *j* is 1. For queries for which *k'*<*k* documents are retrieved, the last summation is done up to *k'(Manning, Raghavan, & Schütze, 2008).*

I.5. Conclusion

In this chapter, we have introduced thebasic concepts oftextual informationretrieval through the studyof the IRprocess andIR models. Each of theproposedmodels tried to solveinherent problems ofinformation retrieval. We also discussedthe evaluationof IR systemsthroughthe introduction ofevaluation measures.

Weare now witnessingmore and moreto the proliferation ofstructuredor semi-structured, i.e., XML documents.Given theirnature,mixingcontent andstructure, XMLdocuments provideorbring up to datea number ofproblemsof classicIR. This subject will be detailed innext Chapter.

Chapter II.

XML Information Retrieval

II.1. Introduction

Since their creation, the markup languages make it possible to archive the electronic documents in a structured or semi-structured form. From SGML[6] (Standard Generalised Markup Language) (Standardization, 1986), to the HTML[7] (HypertText Markup Language) which is an adapted exploitation of the SGML to the WWW (World Wide Web). We arrive today at a standard and universal format of data exchange, known under the name of XML[8] (eXtensible Markup Language). XML is a recommendation (February 1998) of the W3C[9] (World Wide Web Consortium) allowing to combine content and structural information. We focus in the context of our research on structured documents, particularly XML documents.

These particular documents, combining content and structural constraints require the development of new IR approaches that better handle these two components. This leads to revisit all stages of the IR process, namely: querying, indexing, ranking, presenting and evaluating (Lalmas, 2009).

In this chapter, we present the various aspects of the XML IR, in particular the most cited approaches in the XML IR literature. This chapter is organised as follows:

- SectionII.2 presents the basic concepts related to the semi-structured documents;
- SectionII.3 described the specific issues of the XML IR, two classes of approaches are to be distinguished: database oriented approaches, and IR oriented approaches;
- SectionII.4presents the various techniques of indexing of the XML documents;
- SectionII.5described an outline of the XML query languages,taking into account the structural aspect of these documents;
- SectionII.6 presents the various XML retrieval models suggested in the literature;
- SectionII.7represents some examples ofXML information retrieval systems, in fact, Hyrex, TIJAH and XFIRM;
- SectionII.8describes the various concepts related to INEX EvaluationCampaign.

---

[6] SGML : http://www.w3.org/MarkUp/SGML/

[7] HTML : http://www.w3.org/TR/html/

[8] XML : http://www.w3.org/XML/

[9] W3C : http://www.w3.org/

II.2. Basic XML Concepts

II.2.1. XML Documents

XML (eXtensible Markup Language) is to some extent an improved HTML language making it possible to define new tags. It is indeed a language allowingstructuring the documents thanks to tags.

Contrary to HTML, which is considered as a definite language (with a limited number of tags), XML can be regarded as a metalanguage allowing defining other languages, i.e., to define new tags. XML is a subset of SGML, defined by standard ISO8879 in 1986(Standardization, 1986), used in the field of the EDM (Electronic Document Management). XML takes the main part of the functionalities of SGML. It is thus about a simplification of SGML.

The force of XML lies in its capacity at being able to describe any field of data thanks to its extensibility. It makes it possible to structure;to establish the vocabulary and the syntax of the contained data. Actually, XML tags describe the contents rather than the presentation (contrary to HTML). Thus, XML makes it possible to separate the contents from the presentation. XML was developed by the XML Working Group under the support of the W3C since 1996(Consortium, 1998). Since 10 February 1998, the specifications XML-BASED 1.0 were recognised like recommendations by the W3C, which in fact a recognised language. All related document to the XML standard is consultable and downloadable on the Web site of the W3C (Consortium, 1998).

II.2.2. Notion of Structure

The logical structure of an XML document is defined by tags enclosing the portions of information. A tag (or label) is a succession of characters framed by "<" and ">", such as for example "<tag_name>". Logical components of an XML document are called XML elements. An element is an identified semantic unit, delimited by a begin tag and an end tag, such as for example: "<my_tag>" text "</my_tag>". The elements can be overlapping:

```
<Document type= "Book">
       <Title> Introduction to modern information retrieval </Title>
       <Author>
              <Name>Gerard Salton</Name>
       </Auteur>
       <Year>1984</Year>
</Document>
```

Figure II.1 : Example of an XML document

The attributes of XML element are specified at the beginning of the element and after the name of the tag, by using the following syntax: *attribute_name= attribute_value*. For instance:

<my_tag attribute_name = attribute_value>text </my_tag>.

II.2.2.1. Structure ofXML documents

XML provides a way to check the syntax of a document thanks to the DTD (*Document type definition*). A DTD is a file describing the vocabulary (tag names and their attributes) and logical structure (syntax) of XML documents. Thus a XML document must follow scrupulously the conventions of XML notation and can possibly refer to a DTD describing the possible overlap of the elements. A document following the rules of XML is calledwell-formed document. An XML document having a DTD and being in conformity with this one is called valid document.

II.2.2.2. Decoding an XML document

XMLallowsthe definition of an interchange format according to the user needs and offers mechanisms to check the validity of the produced document. It is thus essential for the user of an XML document to be able to extract the data of the document. This operation is possible using a tool called parser. The parser allows on the one hand extracting the data of a XML document and on the other hand checking the validity of the document.

II.2.3. Advantages of XML

The Extensible Markup Language has several advantages (Saint Laurent & Petitjean, 2000), among which we mentioned the following:

- *Legibility*: theoretically no knowledge must be necessary to include/understand the contents of an XML document;
- *A tree structure*: allowing to model the majority of the data-processing problems;
- *Universality and portability*: the various character sets are taken into account;
- *Spreadable*: it can be easily distributed by any protocol (example: HTTP);
- *Integrability*: an XML document is usable by any application containing an XML parser;
- *Extensibility*: an XML document must be able to be usable in all the fields. Thus, XML is particularly adapted to the exchange of information and documents.

II.2.4. XML Standards

II.2.4.1. DOM

DOM (Document Object Model) (Wood et al., 1998) is an application programming interface (API) giving access to the structure and the contents of an XML document. This API makes it possible to generate the tree of objectscorresponding to an XML document.These objects can be XML elements or theirs attributeshave values which are contained on the leaf nodes. Therefore, a DOM tree is composed of: root node; attributes; internal nodes; and leaf nodes. For instance, in the following figure we consider a sample XML document:

```
<Document>
      <Body>
            <Section>
                  <Title> Information retrieval </Title>
                  <Parag> Information retrieval is …. </Parag>
            </Section>
            <Section>
                  <Title> Artificial intelligence </Title>
                  <Parag> This field is …. </Parag>
            </Section>
      </Body>
</Document>
```

Figure II.2: SampleXML document

According to DOM structure, the XML document of Figure II.2is represented as follows (seeFigure II.3):

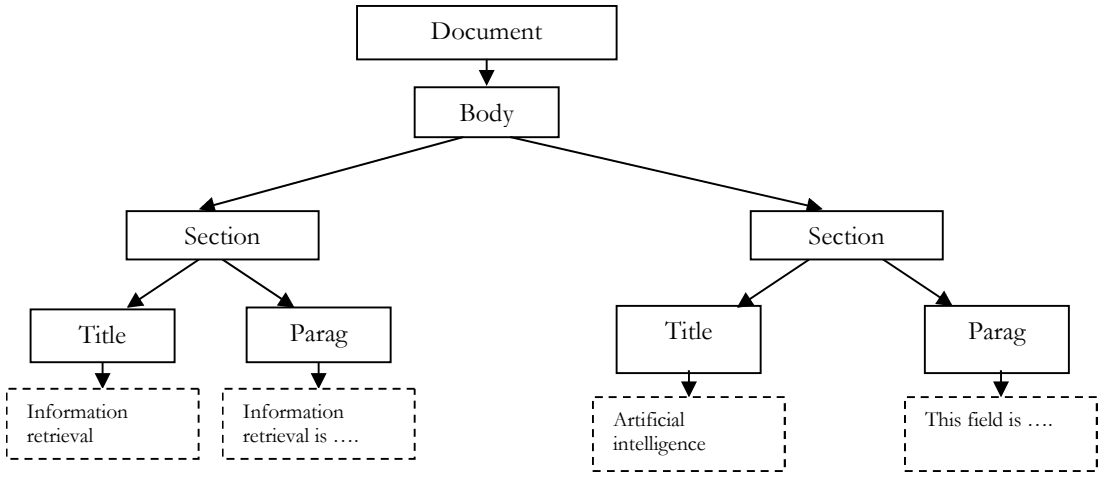


Figure II.3: DOM representation of the XML document of Figure II.2

II.2.4.2. XPath

XPath (*XML Path language*) is a recommendation of the W3C to query XML documents(Clark & DeRose, 1999). Its principal purpose is to access or navigate through the various parts of XML documents by allowing the addressing via location paths. A location path provides the instructions making it possible to access a precise location of the XML document. XPath uses a compact syntax (not-XML) to facilitate its use in URI (*Uniform Resource Identifier*). XPath acts on the logical structures of an XML document, rather than on its apparent syntax. The name XPath comes from the use of the function "access paths", like the URL (*Uniform Resource Locator*), to navigate inside the hierarchical structure of a XML document(Clark & DeRose, 1999).

For the path expressions, the concept of axis is used. Axes allownavigating to a precise location in the DOM structure of the XML document. The following axes are available:

- The child axis: contains the direct descendent of the contextual node;
- The descendent axis: contains the descendants of the contextual node; it never contains nodes of the type: attribute or namespace;
- The parent axis: contains the parent of the contextualnode, if it exists;
- The ancestor axis: contains the ancestors of the contextual node.This axis always contains the root node, except if the contextual node is itself the root;
- The following axis (following-sibling) contains all the target nodes of the contextual node; if that one is an attribute or a namespace, the targeted following is empty;
- The preceding axis (preceding-sibling) contains all the predecessors of the contextual node; if the contextual node is an attribute or a namespace, the preceding target is empty;
- The attribute axis: contains the attributes of the contextual node; the axis is empty when the node is not an XML element;
- The namespace axis: contains the namespaces nodes of the contextual node; the axis is empty when the contextual node is not an XML element;
- The self-axis: contains only the contextual node;
- The descendant-or-self axis: contains the contextual node and its descendants;
- The ancestor-or-selfaxis: contains the contextual node and its ancestors. This axis will always contain the root node (Clark & DeRose, 1999).

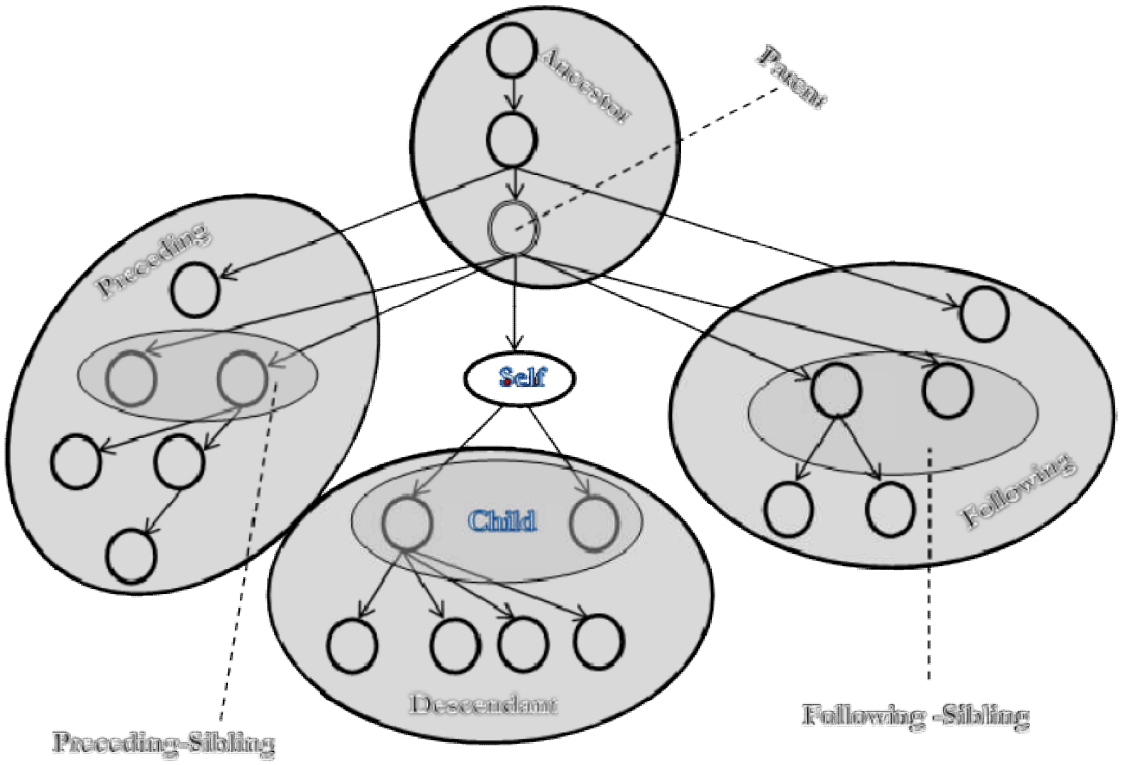


Figure II.4 : Examples of Axes in XPath DOM representation(Lalmas, 2009)

II.2.4.3. XQuery

Since October 1999, the W3C consortium works on the problem of interrogation of XML documents. The fruit of the efforts of the consortium is the XML Query Language or XQuery. This language was conceived to make it possible to create precise queries while being able to adapt to any type of source of XML data, i.e., databases, XML documents or others.

XQuery can be used with XML documents validated by XML schemas, DTD or simply well-formed XML documents. XQuery is a language based on the expressions of the type FLWR (*for, let, where, return*). AnXQuery script (or program) will always contain one or more expressions and in optional functions and definitions.

a) Path Expressions

The path expressions resemble much to those of the XPath language. Let take for example the XML document of Figure II.5, in which the attribute "num_pers" represents a number associated with a student and the value with the note with this one.

```
<Exam>
      <scorenum_pers="001">80</score>
      <score num_pers="012">75</score>
      <score num_pers="525">99</score>
      <score num_pers="601">60</score>
</Exam>
```

Figure II.5 : Sample XML document

The following path expression makes it possible to return the text contained in the node whose value of the attribute "num_pers" is equal to that of the variable $a:

//exam/score [@num_pers=$a] /text ()

b) FLWR Expressions

This syntax is used in various XML query languages. The name comes from:*For, Let, Where and Return*.

- *For*: provides a mechanism of iteration;
- *Let*: allows the assignment of variable;
- *Where*: the clauses *For* and *Let* generate a set of nodes which can be filtered by one or more predicates in a clause where;
- *Return*: generate the result of FLWR expression.

We mention here a simple example of a FLWR query. The purpose of this query is to present a comparison of the prices of the similar books (having the same title) where author is Mohamed Dib, in two bookshops displaying their products on the web.

```
<Books>
{for   $a in document("books.xml")//books/book[auteur='Mohamed Dib'],
       $b in document("products.xml")//products/book[@author='Mohamed Dib']
       where $a/title = $b/title
       return
              <book>
                     <price1>{$a}</price1>
                     <price2>{$b}</price2>
              <book>
}
</Books>
```

Figure II.6: Example of XQuery topic

II.2.5. OtherXML Formats

The interest to have a common format of information exchange depends on the professional context in which the users intervene. This is why many formats of data derived from XML appear (there is more than one hundred):

- OFX: Open Financial eXchange, for the information exchanges in the financial world;
- MathML: Mathematical Markup Language, allowing representation of mathematical formulas;
- CML: Chemical Markup Language, for the description of the chemical compounds;
- SMIL: Synchronised Multimedia Language Integration,allows the creationof multimedia presentations by synchronising various sources: audio, video, text, etc.

II.3. XML IR Challenges

II.3.1. Retrieved Information Granularity

In traditional IR,retrieval results (answers)are returnedto the users in the form of a list of documents. Thesedocuments can contain heterogeneous contents. Thisdemands to users to scroll documents to locatethe desired information within the document.

Structured IRoffers to the user the possibility to have answers in a more significant form. These answers are likely to be auto-explanatory and precise units of information. That means that the contained information in an answer (retrieved unit) does not depend on another to be understood. The units of information are subtrees (XML elements) of the XML documents.

The main issue here is how to select the “right” unit, not too long (with noisy information) and not short (not understandable).

The evaluation of the relevance of an XML element with respect to a query is therefore done according to two dimensions, which are: exhaustivity and specificity.

Exhaustivity describes up to which degree the element discusses the subject of the query. Whereas, specificity describes up to which degree the element is focused on the subject of the query.

The retrieval principle in structured documents can be defined by: "A system should always find information unit the most exhaustive and specific answering a query".

II.3.2. Specific Issues of XML IR

Structural dimension of XML documents raises several issues relative to each step of the IR process.

The first specific issue is the document indexingstep. Indexing XML document must take into account the structural information inaddition to the content. As for this phase several questions arise: What structural information of the documents must be indexed? How to associate this structure to the content? How to evaluate the index terms relevance, i.e., how to evaluate the importance of a term within the element, the document and the collection?

The second issueisrelative to the querying of the structured documents corpora. AnXML IRsystem must be able to make possible to a user to express his information need in a simple way and by exploiting both types of contained information (textual and structural).

The last issueis that of the ranking models of the retrievedunits. The retrieval system must be able to decide the granularity of information unit to retrieve if the query is"Content Only" (CO), i.e.composed only of terms. If it is about a "Content And Structure" (CAS) query, i.e.contained textual and structural constraints.

- the user specifies the type of elements to be retrieved;
- The user does not specify the type of elements to be retrieved.

In the second case, it is the task of the system to decide granularity (XML element) of information unit to retrieve.

II.3.3. XML IR Approaches

The XMLIR includes two approaches whichattempt to propose methods for: indexing, querying, retrievingand ranking of the XML documents. These two approaches are:

- Data-Centric Approaches based on techniques developed by the databases community, and they see the XML documents as collections of data, typified and relatively homogeneous.
- Document-Centric Approachesaredeveloped by IR community, and are focused on applications considering the XML documents in a traditional way, i.e., the tags are only used to describe the logical structure of the documents.

Table II.1 summarises some specific points to each approach for each phase of the retrieval process.

	Data-Centric Approaches	Document-Centric Approaches
Indexing	- Confuse the concepts of indexing and storage: all textual and structural information of the documents are stored within tables in databases.	- Use of Traditional techniques for the extraction of the index terms. - New issues are raised concerning the structure: What must be indexed of

	- This poses a problem for CO retrieval, since the textual contents are indexed as strings. - Propose optimal schemes of storage for the structure of the documents.	the structure of XML documents? How to associate this structure to the content of document?
Query languages (Interrogation)	- Historically, data-centric approaches are the first to propose query languages for the interrogation of XML documents. - These query languages are based on syntaxes close to SQL, and make it possible the user to express precisely the structural constrains. - Queries must always relate to well defined structural constraints. The user must moreover specify the type of element to be retrieved by the system, even if he does not have a precise idea on the question.	- Try to simplify these query languages with regard to the structural constraints. - New functionalities concerning textual content retrieval are proposed (use of the predicate 'about' instead of 'contains', or even of Boolean operators in the textual content constraints).
Query Processing	- Evaluate in an exact way *attribute=value* expressions type. - Query processing is carried out in a Boolean way and it is not possible to return a sorted list of results to the user.	- Evaluate the degree of relevance between the query and the information units by assigning a relevance score. - The interest is double: first of all, select the informationunits that best answering the user's need; secondly, propose to him a sorted list of retrieval results.

Table II.1: Characteristics of the two approaches for each phase of the retrieval process

The suggested solutions by the IR community can be used as "*top layer*" with databases oriented solutions. This "*top layer*"is primarily used to integrate the concept of relevance in the retrieval process, by complementing the approaches proposed by the DB community for the storage and the interrogation of the documents.

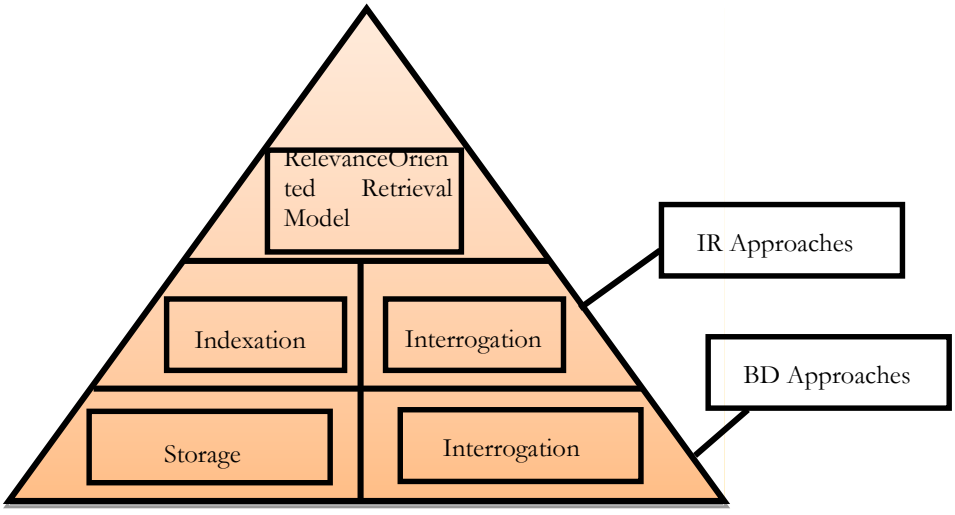


Figure II.7: Fields of competence of the DB and the IR(Sauvagnat, 2005)

II.4. Indexing of XML documents

The indexing of the XML documents differs from that used for "flat" (textual) documents because of the structural dimension which is added to the content. In this context several questions arise: What structural information of the documents must be indexed? How to associate this structure to the content of XML documents? How to evaluate relevance of the indexed terms, i.e., how to evaluate the importance of a term within the element, the document and the collection? We try to answer each one of these questions in the following sections.

II.4.1. What should beindexed?

As aforementioned, structured documents comprise two types of information: textual information (content) and structural information (the structure). At first sight, we think that the simplest manner to index these documents is to consider only textual information, which makes this process similar to that used in traditional IR. Two weaknesses of this manner of indexing can be cited: firstly, no search on the structure is possible; secondly, the granularity of information used remains always the entire document.

According to Sauvagnat (Sauvagnat, 2005), several aspects must be covered by the indexing schema:

- Allowing the rebuilding of the XML document fragmented up in the storage structures;

- Allowing the processing of the path expressions on the XML structure;

- Allowing accelerate navigation inside XML documents;

- Authorising the processing of vague and precise predicates on the contents of XML documents;

- Allowing information retrieval by using keywords.

Indexing approaches of the structured documents are characterised by two dimensions:

- The storage schema of the documents;
- Possible types of transformations between the XML documents and the structures of storage.

Concerning the storage scheme, two types of approaches can be mentioned:

- DBMS oriented approaches (middleware of transformation);
- XML based storage models (native) which store complete documents or parts of documents.

For the possible transformation type's dimension, wedistinguish:

- Model based transformation approaches: these approaches create a generic databasescheme which reflects the data model of the XML format. These approaches are regarded as extensible and do not need the DTD of XML documents;
- Structure based transformation approaches: these approaches build an indexscheme which takes into account specificities of the application.

In (Sauvagnat, 2005), author adopted a classification which relates to the type of information, this classification allows to better include/understand issues raised by each indexing type.

II.4.2. IndexingTextual Information

The problems of indexing textual content remains of topicality in the field of structured IR. Databases oriented approaches considerleafnodes as being the textual unit of indexing; whereas IR oriented approaches consider the term. Two problems prove to be interesting to detail in indexing of the textual content: firstly;term scope and secondly the problem of their weighting.

II.4.2.1. Scope ofIndexing Terms

We attempt in this part to answer to the question of: *how to associate the terms to structural information?* Two solutions were proposed in the literature(Sauvagnat, 2005):

- The "Nested subtrees" indexing approach which seeks to incorporate the contents of the nodes;
- The "disjoined units" indexing approach which indexes all the contents of the nodes separately.

a) Nested Subtrees

This category of approachesconsiders thatthe full text ofeach node of theindex schemeas an atomicdocument and thereforepropagatestheleaf nodes contentin thedocument tree.

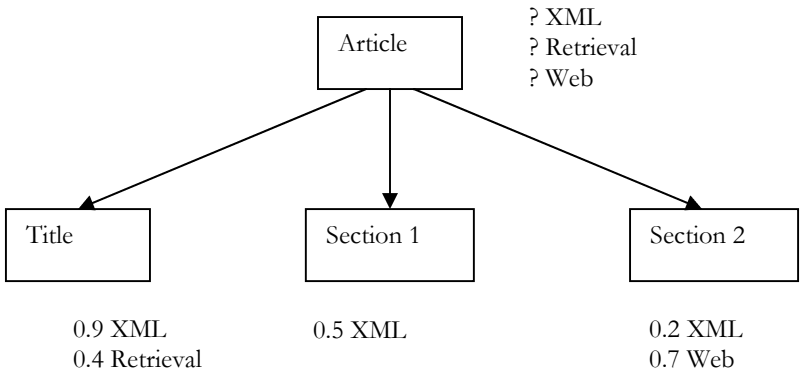


Figure II.8: Nested subtrees approach example

b) Disjoint Units

In these approaches the XML document is decomposed into disjoined units, in such way that the text of each XML node of the index is the union of one or of more than these disjoined parts.

II.4.2.2. Weighting ofIndex Terms

In the case of DB oriented approaches, often no term weighting is carried out because the text of the leaf nodes is considered as being only one unit. Contrarily, inIR oriented approaches term weighting is of a great importance. Term weighting must be seen in another way,new formulas making it possible to evaluate the importance of terms within the element, the document and the collection are more suitable, ***idf*** (*Inverse Document Frequency*) used in traditional IR was adapted for the structured IR under the name of ***ief*** (*Inverse Element Frequency*). Thus the ***tf.idf***formula was replaced by ***tf.itdf*** (*Term Frequency-Inverse Tag and Document Frequency*), which makes it possible to computethe relevance of a term ***t***of the XML element***e***within a document ***d***.

Several other parameters can be taken into account in addition to the term frequency in the XML element. We mentionelement length, the average length of the elements in the document, collection, etc.

II.4.3. Indexingthe Structural Information

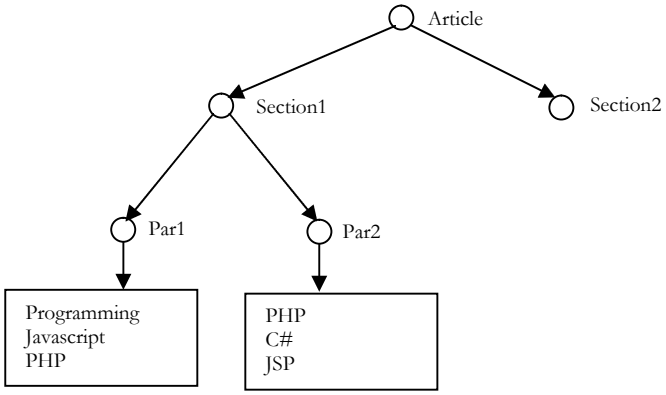
We distinguish in the literature three classes of approaches for indexing structural information, the first known as fields based approaches, the second known as paths based approaches, and the last known as trees based approaches. These three classes of approaches are independent of the manner of using textual information (IR or DB approaches).

II.4.3.1. IndexingBased on Fields

In this method the document is represented as a set of fields and content associated with each field. Several ways can be used to obtain the various fields of an XML document:

- they can be coded as metadata in the XML files, i.e., by using RDF;
- in the case of a transformed document format into XML, they are extracted from the document in its original format;
- they can be found using various extraction techniques;
- they are simply extracted from the DTD or the associated XML schema.

The index is built by combining the name of the field with the terms of the content to allow a restricted retrieval upon certain fields.

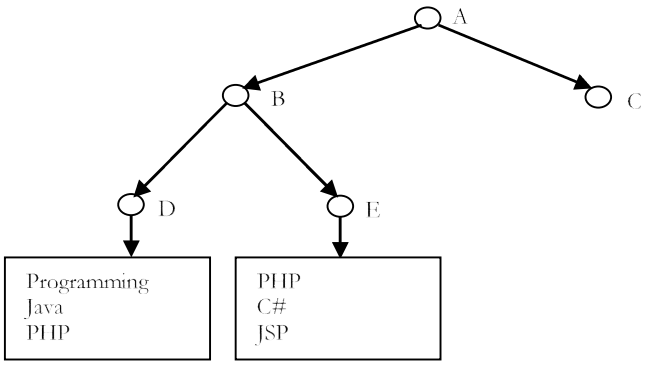


Terms	Fields
Programming	(Par1)
Javascript	(Par1)
PHP	(Par1), (Par2)
C#	(Par2)
JSP	(Par2)

Figure II.9: Example of fields based indexing

II.4.3.2. IndexingBased on Paths

This type of techniques facilitates navigation inside the documents by allowing the resolution of the XPath expressions; it also allows finding documents having known values for certain elements or attributes, and uses full text index on the content. These techniques suffer from the difficulty in finding the relations ancestor-descendent between the various nodes of the documents. The following figurerepresents an illustration of this type of indexing.



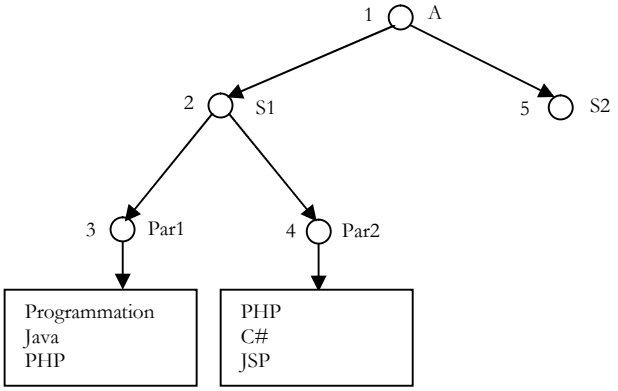
Terms	Paths
Programming	(/A/B/D)
Java	(/A/B/D)
PHP	(/A/B/D), (/A/B/E)
C#	(/A/B/E)
JSP	(/A/B/E)

Paths	Documents
(/A)	Doc1,…
(/A/B)	Doc1,…
(/A/B/D)	Doc1,…
(/A/B/E)	Doc1,…
….	……..

Figure II.10: Example of paths based indexing

II.4.3.3. IndexingBased on Trees

This type of indexing makes it possible to solve the difficulty of the paths based technique (namely: to find the relations ancestor-descendent) because the nodes of the tree are numbered in the tree structure of the documents.



Terms	Nodes
Programming	3
Java	3
PHP	3, 4
C#	4
JSP	4

Figure II.11: Example of trees based indexing

Several techniques belong to this class, among them:

- ANOR index(Y. K. Lee, Yoo, Yoon, & Berra, 1996);
- EDGE approach(Florescu & Kossmann, 1999);
- BINARY approach (Florescu & Kossmann, 1999);
- XPath Accelerator index (Grust, 2002);
- XFIRM approach (Sauvagnat, 2005).

II.5. Query Languages

The query languages make it possible to providea tool of expression of the user need. The structural aspect of the XML documents allows extendingof the expression possibilities, which gave rise to a new form of interrogation. In addition to the Content Oriented (CO) queries, the user can add structural constraints to create queries known as Content and Structure (CAS) queries. XML needs a query language that is as flexible as XML itself. For querying XML documents, the query language needs to be able to preserve order and hierarchyand must be capable of dealing with all the information structures found in the XML Schema specification.

In this section, we consider the principal query languages proposed in the literature. These languages come mainly from the databases community, and their syntax is often derived from SQL. Recently, some IR oriented languages made their appearance (for instance, NEXI, XFIRM).

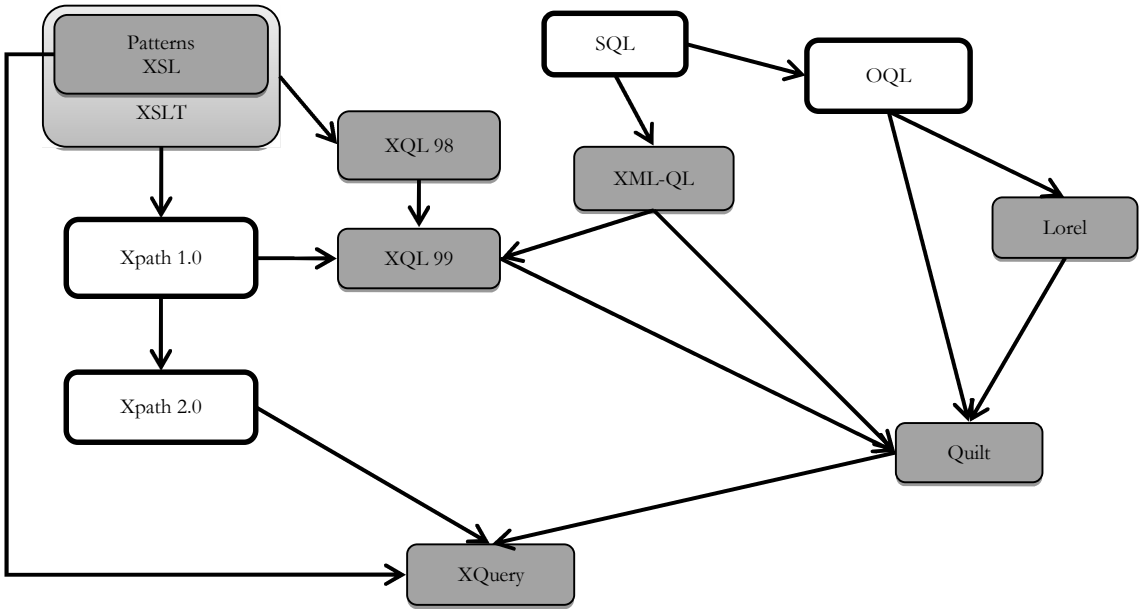


Figure II.12: XML querying languages (Sauvagnat, 2005)

II.5.1. XML-QL

Matching data using patterns, XML-QL (Deutsch, Fernandez, Florescu, Levy, & Suciu, 1998)uses element patterns to match data in an XML document. XML-QL is very different from XPath because it does not use path expressions but patterns to match fragments of one or more XML documents. It is based on a WHERE-IN-CONSTRUCT statements, very similar to the SELECT-FROM-WHERE syntax in SQL. This XML query language is designed with the following features:

- it is declarative, like SQL;
- it is relational, e.g. it can express joins;
- it can be implemented with known database techniques;
- ability to construct completely new XML fragment and return it as a result from the query;

Other interesting features:

- Constructing explicit root element;
- Grouping of data;
- Transforming XML data;
- ability to combine and query data from different data sources;

As an example, to find those authors who have published books for McGraw Hill, the XML-QL query will be:

```
WHERE
<bib><book>
             <publisher><name>McGraw Hill</></>
             <title>$t</>
             <author>$a</>
</book></bib> IN "bib.xml"
   CONSTRUCT <result><title>$t</><author>$a</></>
```

The $t and $a are variables that pick out contents. The output of this XML-QL query is a collection of author names.

The *WHERE* clause contains a pattern that will be matched against the source document. The *IN* clause specifies the source that the query processor will use to match the pattern from the *WHERE* clause. Finally, the CONSTRUCT clause specifies how the query will return results (if any). Some facts of interest:

- The source could be XML file located by any valid URI which makes XML-QL very powerful (XPath and XQL always perform queries on a tree representation of already parsed XML document);
- Several sources could be combined, so that the query will operate on many XML documents (just like in SQL the data source could be formed of many tables or views);
- The source could also be a bound variable, representing some XML fragment of a document.
- XML-QL is able to express joins (just like SQL).

II.5.2. XQL

XQL (*XML Query Language*) (Robie et al., 1999)is a query language for collections of XML documents to which semantics is close to that of XQuery and XPath.It was Submitted to the W3C by WebMethods and Microsoft, XQL remained at the stage of the proposal. The objective of XQL was to remain simple while offering more possibilities than XPath. XQL uses path expressions which most are valid XPath expressions. XQL don't offer axis concept and its expressions exploit the child axis expressions. XQL offers some features that cannot be found in XPath:

- all and any semantics for specifying whether the predicate holds true if all items in a node-set meet the predicate condition or at least one item in the set meets the condition (XPath uses only "any" semantics);
- $intersect$ operator in XQL for intersection of the node-sets generated from two path expressions;
- The subscript operator offers more options than the one in XPath.

II.5.3. Quilt

Quilt[10] is anXML query language created by W3C XML Query Working group members (D. Chamberlin, Robie, & Florescu, 2001) and conceived for the interrogationof heterogeneous

[10]Quilt papers - http://www.almaden.ibm.com/cs/people/chamberlin/quilt.html

data sources.It is flexible to interrogate a wide spectrum of XML information sources. This query languageis inspired from the design ofXPath(Clark & DeRose, 1999), XML-QL(Deutsch et al., 1998), SQL(D. D. Chamberlin & Boyce, 1974)and OQL(Alashqur, Su, & Lam, 1989)by following a strategy to borrow features from these languages that seem to have strengths in specific areas. Members of W3C XML Query Working group took from XPath (Clark & DeRose, 1999) and XQL(Robie et al., 1999) syntax for navigating in hierarchical documents; and took the binding variables notion from XML-QL. From SQL, they tookkeywords based clauses that provide a pattern for restructuring data, for instance, the "SELECT-FROM-WHERE" pattern. Finally, they took from OQL the notion of a functional language.

The input and output of a Quilt query are fragments of XML documents, XML documents, or collections of XML documents. These inputs and outputs can be seenas instances of a data model known as the XML Query Data Model. This data model is a refinement of the data model described in the XPath specification, in which an XML document is modelled as a tree of nodes. Quilt is based on a*FOR-LET-WHERE-RETURN*(FLWR) syntax. A simple query could look like:

```
FOR    $vendor IN document("vendors.xml")/VENDOR_LIST/VENDOR
LET    $ven_phone = $vendor/PHONE
WHERE  $vendor/NAME="ABC"
RETURN count($ven_phone)
```

Few facts of interest for the Quilt syntax are:

- The query uses path expressions to locate nodes (just like in XPath)
- A variable binding could be specified with the help of already bound variable ($ph in our case).
- Variables in Quilt are bound to the whole element and not only to its content (so the result of the query need not be wrapped in a <VENDOR_PHONE> element like in XML-QL)

Other great features of Quilt like quantifiers, filters, *AFTER* and *BEFORE* modifiers, union of queries and views will not be discussed. Kweelt[11] is a Quilt based query engine that follows most of the specification.

II.6. Query Processing

In the context of the structured IR we distinguish two approaches of query processing:

- The databases oriented approach, which process the content of XML documents in a Boolean way;
- The IR oriented approach, which tries to assignrelevance scores to the nodes of the XML documents.

[11] Kweelt - http://db.cis.upenn.edu/Kweelt/

The queries are of two types: content only and contentandstructure. Several models were proposed in order to process these queries. These models are at the base of the traditional IR models adapted to the context of the structured IR. In the following sections, we give an outline on some models suggested in the literature.

II.6.1. Extended Vector Space Model

The Extended Vector Space Model tries to measure the similarity of each element with respect to a query. Most of the approaches use the propagation method and indexes the documents by using"nested subtrees" method, i.e., the terms are propagated in the tree of the XML document. The retrieval result of the query is aranked list of XML elements.

One of the first adaptations of the vector space model to the context of structured documents is that of Fuller etal.(Fuller, Mackie, Sacks-Davis, & Wilkinson, 1993). In their model, the similarity of a node with respect to a queryis expressed by the following formula:

$$Sim(q,n) = \alpha(T).\cos m(q,n) + \sum_{k=1}^{s} \frac{\cos m(q,n_k)}{\beta^{k-1}} \quad \text{(II.1)}$$

Where:

- $\alpha(T)$: factor representing the type of the XML node.
- s :the number ofdescendant nodes$n_k$of$n$.
- β :a parameter which ensures the non-introduction of a bias by the number of descendant nodes.

The*cosm* function is definedas follows:

$$\cos m(q,n) = \sum_{i=1}^{T} \frac{w_i^q * w_i^n}{|n|} \quad \text{(II.2)}$$

Where w_i^q et w_i^n representrespectively theweight of the termt_i in the query q and the XML node n; $|n|$ represents the number of terms in the XML node n.

The relevance of an XML node can be computed and combined with the relevance of the descendant nodes. For the processing of the CAS queries the model can be applied recursively to each subtree of the hierarchy. Thereafter an aggregate score is performed.

II.6.2. Probabilistic Model

The adaptation of theprobabilistic modelto the context ofstructuredIRmust take into accountthe structural aspectofXMLdocuments.Gövert et al.(Gövert, Abolhassani, Fuhr, & Großjohann, 2002)proposed a probabilistic based modelimplementedinthe *Hyrex*search engine. *Hyrex*uses the XIRQL query language.In this modelleafnodesare not indexedand terms arepropagatedto the nearestindexablenodes.The relevance of anodeis computed byspreadingthe weight ofterms uponthe document hierarchy. A factorcalled "*augmentation factor*" is used to reduce theweightof a termduring propagation process.

II.7. Examples of XML IR Systems

We present in this section three examples of XML IR systems: *Hyrex*(Fuhr, Gövert, & Großjohann, 2002), *TIJAH*(Mihajlović, Ramirez, De Vries, Hiemstra, & Blok, 2005), and the *XFIRM*(Sauvagnat, 2005) systems.

II.7.1. Hyrex Retrieval System

The Hyper-media Retrieval Engine for XML (HyREX) provides an implementation of the XIRQL query language.

This retrieval strategy implemented in HyREX in order to process the INEX content-only topics try to retrieve those document components (elements) which answer the information need in the most specific way.

Fuhr et al. have defined the "atomic" units within structured documents. Their definition serves two purposes: first, for relevance-oriented search, where no type of result element is specified, these units are the retrievable units; second, they provide a context within a document which can serve as a meaningful answer to a user's information need; finally, given these units, they can apply for example some kind of *tf·idf* formula for term weighting.

The XIRQL query language is used to query XML documentsincluding structural constraints. XIRQL allow processing of the INEX CAS topics by convertingthem in a fully automatic way to be processedby HyREX. Figure II.13 shows an example of conversion of the topic 24 of INEX.Different elements specific search predicates are appliedin HyREX (eg. phonetic similarity on author names and stemmed search for other query terms).

```
<Title>
       <te>article</te>
       <cw> Jones Smith </cw><ce>au</ce>
       <cw>software  engineering  and  process  improvement</cw>
       <ce>bdy</ce>
</Title>
```

Will be transformed into:

```
.//au//#PCDATA[. $soundex$ "John" $and$ . $soundex$ "Smith"]
```

Figure II.13 : INEX Topics conversion to XIRQL syntax

II.7.2. TIJAH Retrieval System

II.7.2.1. Definition

TIJAH(Mihajlović et al., 2005)is an information retrievalsystem in collections of XML documentsbased on the MonetDBdatabasesystem.It isa system designedinlayers (or levels), which gives ittheappearance of beinga relational DBMS. TIJAHwas developed bya teamconsisting of:JohanList,VojkanMihajlovic, GeorginaRamirez,ThijsWesterveld, DjoerdHiemstra, Henk ErnstBlok, ArjenP.de Vries. Their firstparticipation in theINEXcampaignwasin 2003.Thenin thecampaignof2004 and2005INEXTIJAHsystem wasequipped with aset of techniques.

II.7.2.2. Architecture ofTIJAH System

TIJAH System is consisted of the three traditional levels of a DBMS which are: conceptual, logic and physic levels (see Figure II.14). The designers of TIJAH system carried out some modifications on this architecture in order to establish the link between DBMS and IR systems.

a) Conceptual Level

TIJAH System uses NEXI language as its conceptual level (Trotman & Sigurbjörnsson, 2005). NEXI Expressions are coded in a form similar to the original query. This form is called internal representation.

b) Logic Level

This level is based on the score region algebra. The basic idea of the region algebra based approaches is to present the text in the form of a set of regions, where each region is defined by a position of beginning and a position of end(Mihajlović et al., 2005).

c) Physic Level

TIJAH System uses, in the physical level, the MonetDB. The last stage of the logic level is the translation towards a MIL (*Monet Interpreter Language*) query. The MIL query are carried out by using the primitives MIL which define handling on BATs (Monet Binary Tables).

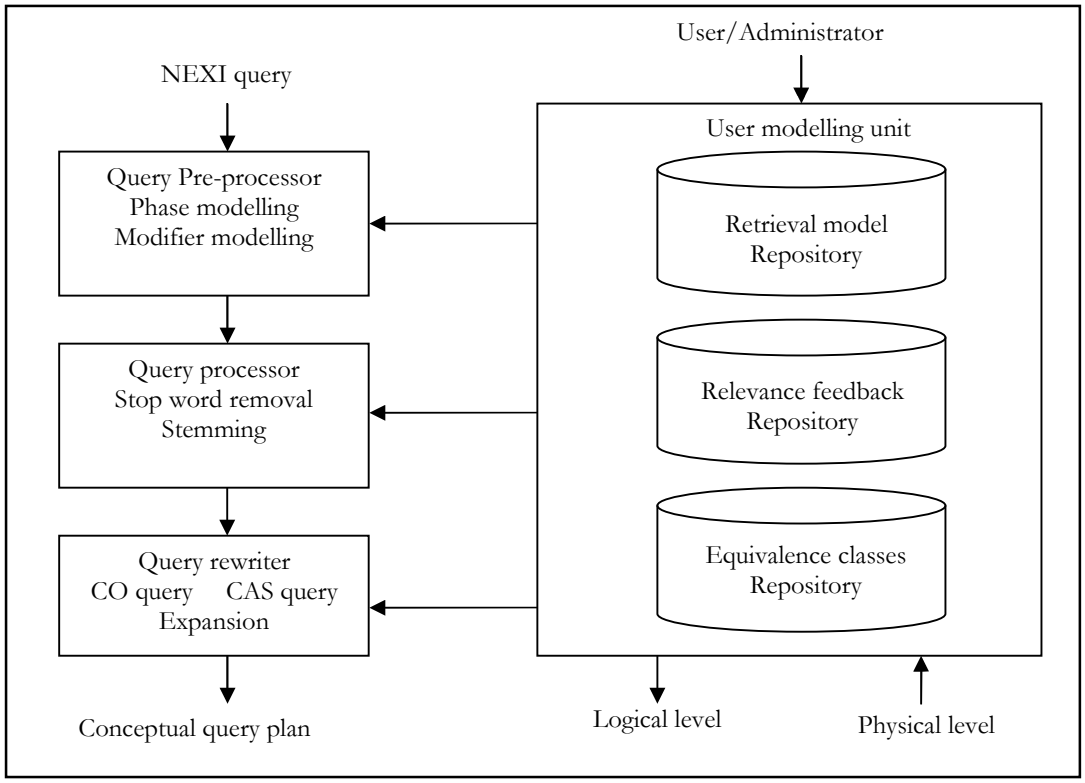


Figure II.14: Conceptual Level of the TIJAH retrieval system(Mihajlović et al., 2005)

II.7.3. XFIRM System

II.7.3.1. Definition

XFIRM System (Sauvagnat, 2005)is based on the XFIRM model (XML Flexible Information Retrieval Model). This model was proposed in order to answer some limitations of other XML IR approaches. It uses a technique of propagation of the relevance of the nodes to evaluatequeries. XFIRM defines: a model for the representation of XML documents, a query language and a process for the evaluation of queries.

II.7.3.2. Documents Representation Model

The model of representation proposed makes it possible to navigate in the tree structure of the XML documents and to represent the content of this structure. It is about a simplification of the XPATH model (Clark & DeRose, 1999). In XFIRM, a structured document *ds* is a tree defined by the units: *NR*, *F*, *A* and *L*.

$$ds = (N, F, A, L)$$

Where:

- N = {n_1, n_2,…} represents the set of internal nodes ;
- F = {nf_1, nf_2,…} represents the set of leaf nodes ;
- A = {a_1, a_2,…} represents the set of attributes ;
- L represents the set of hierarchical links between nodes;

This representation makes it possible to manage effectively heterogeneous documentation. Textual information is located at the level of the leaf nodes. A leaf node is defined by:

$$nf_i = \{(t_1, w_1^i), (t_2, w_2^i), ...\} = \{(t_j, w_j^i)\} \quad \text{(II.3)}$$

Where: $t_j$is a term and $w_j^i$is the term weight in the node *i*

Several weighting formulas can be used in the XFIRM system. In (Sauvagnat & Boughanem, 2006), we can find a study on the impact of the various formulas used for the weight computation of the index terms. These formulas are combinations of the functions *tf* (term frequency), *idf* (inverse document frequency) and *ief* (inverse element frequency).

II.7.3.3. Query Language

According to (Sauvagnat, 2005), the XFIRM query language is characterised by:

- simple syntax which can be seen like a simplification of XPath;
- formulation of queries containing simple key words;
- possibility of formulating constraints on the structure of the documents;
- possibility of formulating more complex queries, while introducing the concept of hierarchy between the various structural constraints;

- possibility of extending the queries thanks to a dictionary of tag names of the various nodes in the corpus.
- expression of the user need can be carried out with XFIRM query language according to four degrees of accuracy:
    - queries with simple key words (P1 queries);
    - queries with conditions on the structure of the XML documents (P2 queries);
    - queries with conditions on the structure with addition of the concept of hierarchy between the various structural constraints (P3 queries);
    - querieswith specification of the unit of information which the user wishes to see retrieved (P4 queries).

II.7.3.4. Query Processing

In XFIRM retrieval system two types of evaluation exist. These evaluation types correspond to each type of queries (CO and CAS).

The evaluation of the CO queriesis carried out in two stages: Firstly, evaluate the similarity between the leaf nodes of the index and the query. Secondly, search relevant and informative subtrees. This is carried out by:

- propagation to the top of the score of the leaf nodes in the tree of the document;
- propagation to the bottom of the score of the document (contextual relevance).

The similarity function for the evaluation of CO queriesis defined as follows:

$$RSV_m(q, nf) = \sum_{i=1}^{T} w_i^q * w_i^{nf} \quad \text{(II.4)}$$

Where: w_i^q represents the weight of the term *i* in the query and w_i^{nf} represents the weight of the term *i* in the leaf node *nf*. Relevance $\boldsymbol{P_n}$ of the node $\boldsymbol{n}$ is defined by the following formula:

$$P_n = \rho * \left|F_n^p\right| . \sum_{nf_k \in F_n} \alpha^{dist(n, nf_k)-1} * \beta(nf_k) * RSV(q, nf_k) + (1-\rho) * P_{racine} \quad \text{(II.5)}$$

Where: $\boldsymbol{F_n}$:is the set of leaf nodes descendants of $\boldsymbol{n}$;

$\boldsymbol{F}$:is the set of the XML document leaf nodes;

And $\beta(nf_k)$: represents a parameter introducing the length of the leaf node.

With regard to the evaluation of content and structure queries, only one processing is carried out. Indeed, P3 and P4 queries will be broken up into queries of the P2 type:

P3=//$P2_1$//$P2_2$//…//$P2_n$

$P4=//P2_1//P2_2//\ldots//ec :P2_i//\ldots//P2_n$

Thus, P3 and P4 queries will be represented by a set of elementary queries. The evaluation is carried out in three stages:

- evaluation of elementary sub-queries;
- evaluation of the P2 queries by taking into account the results obtained by elementary sub-queries.
- evaluationof the structural constraints specified in the query by using the results of the P2 queries.

II.8. INEX Evaluation Campaign

The evaluation of IR systems is a very important phase to compare their performances. Such as TREC (Text Retrieval Conference) (D. Harman, 1993) for the full text IR, INEX (Gövert & Kazai, 2002) is regarded as being the only evaluation campaign (since 2002 to date) of the IR systems in collections of XML documents. The evaluation of the effectiveness of the XML IR systems requires a test collection. A test collection generally consists of a set of XML documents, user requests (topics) and relevance assessments. This test collection is provided to the various participants to make an evaluation of their systems.

II.8.1. Test Collection

The test collections in traditional IR are composed of three parts: a set of documents, a set of users' needs expressed in the form of queries (Topics) and a set of relevance assessments of represented by a list of the relevant documents corresponding to each query. In XML IR, the test collection differs on several points from that used in traditional IR. Although it always consists of the same parts, the nature of these parts is fundamentally different. The XML elements are considered as the atomic retrieval unit. In addition to the keywords, XML queries can contain structure constraints. Consequently the relevance assessments must take into account the structural nature of the XML documents.

II.8.1.1. INEX 2002 Test Collection

The INEX 2002 test collection consists of the full texts of 12107 articles from 12 magazines and 6 transactions of the IEEE Computer Society's publications, covering the period between 1995 and 2002 (494 Mbytes). This collection has an appropriately complex XML structure compared with TREC (192 different elements in DTD) containing scientific articles of varying length.

II.8.1.2. INEX 2005 Test Collection

In 2005, the INEX test collection of documents has been extended with further publications provided by the IEEE Computer Society. A total of 4712 new articles published between 2002 and 2004 have been added to the previous collection of 12107 articles, giving a total of 16 819 articles. The INEX 2005 ad-hoc test collection grew by 228 Mbytes in size.

In addition to the original collection, another collection of XML documents has been added in 2005 specifically for the Multimedia task. This collection is based on "The Lonely Planet WorldGuide" which consists of 462 XML documents (Van Zwol, Kazai, & Lalmas, 2006).

II.8.1.3. INEX 2007 Test Collection

In 2006, Denoyer and Gallinari (Denoyer & Gallinari, 2007) have created a corpus of XML documents based on part of the free encyclopaedia: Wikipedia (Wikipedia). The complete corpus was composed of 8 main collections corresponding to 8 different languages: English, French, German, Dutch, Spanish, Chinese, Arabian and Japanese. Each collection is a set of XML documents built using Wikipedia and encoded in UTF-8. In addition to these 8 collections, they also provided different additional collections for other IR/Machine Learning tasks like categorization and clustering, NLP, machine translation, multimedia IR, entity search, etc.

The INEX 2007 Wikipedia XML English corpus used at the INEX evaluation initiative contains about 659,388 XML documents in English language, densely hyperlinked. On average an article contains 161 XML nodes, where the average depth of a node in the XML tree of the document is 6.72 (Fuhr, Kamps, Lalmas, Malik, & Trotman, 2008).

Denoyer and Gallinari (Denoyer & Gallinari, 2007)introduced different types of tags to represent the fragments of a Wikipedia XML document. These tags are of two types:

- General tags (*article,section, paragraph ...)* that do not depend on the language of the collection. These tags correspond to the structural information contained in the wikitext format (Denoyer & Gallinari, 2007).
- Template tags (*template infobox,...)* represent the information contained into the Wikipedia templates. Wikipedia templates are used to represent a repetitive type of information. For example, each country described into Wikipedia starts with a table containing its population, language, size, etc. The template tags depend on the language of the collection because the templates are not the same depending on the language of the Wikipedia collection used (Denoyer & Gallinari, 2007).

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
       <name id="40774">Base communications</name>
       <conversionwarning>0</conversionwarning>
<body>
       <template name="move to wiktionary"></template>
       <emph3>Base communications </emph3> (basecom):
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple"
       xlink:href="40914.xml"> Communications </collectionlink> services, such as the
       installation, <unknownlink src="operation">operation</unknownlink>,
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple"
       xlink:href="91191.xml">maintenance</collectionlink>, augmentation, modification, and
       rehabilitation of communications networks, systems, facilities, and equipment,
       including off- post extensions, provided for the operation of a military post, camp,
       installation, station, or activity.<emph2>Synonym</emph2>
       <emph3>communications
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink"        xlink:type="simple"
xlink:href="510114.xml">base station        </collectionlink>
       </emph3>.
       <p>Source: from <collectionlink      xmlns:xlink="http://www.w3.org/1999/xlink"
       xlink:type="simple" xlink:href="37310.xml">Federal Standard
       1037C</collectionlink>
       </p>
</body>
</article>
```

Figure II.15 : Example of INEX 2007 Wikipedia XML document (file "40774.xml")

II.8.1.4. INEX 2009 Test Collection

The INEX 2009 test collection comprises 2,666,190 XML documents (a total uncompressed size of 50.7 Gb) and 115 topics (Geva et al., 2010). Starting in 2009, a new document collection based on the Wikipedia has been used. Wikipedia original syntax has been converted into XML format, using both general structural tags ("*article*", "*section*", "*paragraph*", etc.), typographical tags (*emphatic*, *italic*, *bold*, etc.), and frequently occurring link-tags (Geva, Kamps, & Trotman, 2009). The annotation used has been enhanced with semantic markup of articles and outgoing links, based on the semantic knowledge base YAGO, explicitly labeling more than 5,800 classes of entities like "*persons*", "*movies*", "*cities*", etc. The collection contains 101,917,424 XML elements of at least 50 characters (excluding white-space).

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- generated by CLiX/Wiki2XML [MPI-Inf, MMCI@UdS] $LastChangedRevision: 92 $ on 16.04.2009
15:24:05[mciao0825] -->
<!DOCTYPE article SYSTEM "../article.dtd">
<article xmlns:xlink="http://www.w3.org/1999/xlink">
<header>
       <title>Default</title>
       <id>8000</id>
       <revision>
              <id>242931647</id><timestamp>2008-10-04T09:48:59Z</timestamp>
              <contributor>
                     <username>Cyfal</username><id>4637213</id>
              </contributor>
       </revision>
       <categories>
              <category>All disambiguation pages</category>
              <category>Disambiguation pages</category>
       </categories>
</header>
<bdy>
<p><b>default</b>,as in failing to meet an obligation, may refer to:
<list>
       <entry           level="1"           type="bullet"><link           xlink:type="simple"
       xlink:href="../gan/Byron_C$enter=2C_M$ichigan.xml">Default (law)</link>
       </entry>
       <entry level="1" type="bullet">
              <link xlink:type="simple" xlink:href="../838/58838.xml">
              Default (finance)</link></entry>
</list>
</p>
<p><b>default</b>,as a result when no action is taken, may refer to:
<list>
       <entry level="1" type="bullet"><information wordnetid="105816287" confidence="0.8">
              <datum wordnetid="105816622" confidence="0.8"><link xlink:type="simple"
              xlink:href="../316/957316.xml">Default (computer science)</link></datum>
              </information>—also contains consumer electronics usage
       </entry>
       <entry level="1" type="bullet">
              <link xlink:type="simple" xlink:href="../639/889639.xml">Default logic</link>
       </entry>
</list>
</p>
<p>It may also refer to:
<list>
       <entry level="1" type="bullet">
              <musical_organization wordnetid="108246613" confidence="0.8">
              <group wordnetid="100031264" confidence="0.8">
              <link xlink:type="simple" xlink:href="../344/9159344.xml">Default
       (band)</link></group>
              </musical_organization>,a Canadian post-grunge and alternativerock
       band
       </entry>
       <entry level="1" type="bullet">
              <link xlink:type="simple" xlink:href="../734/3841734.xml">defaults
       (software)</link>,a command line utility for plist (preference) files
       </entry>
</list>
</p>
<p>
```

```
<table style="background:none">
<row>
	<col style="vertical-align:middle;">
		<image width="30px" src="Disambig_gray.svg"></image>
	</col>
	<col style="vertical-align:middle;">
		<it>This  page lists articles associated with the same title. If an<weblink
		xlink:type="simple"
	xlink:href="http://localhost:18088/wiki/index.php?title=Special:Whatlinkshere/Default&amp;namespace=0">internallink</weblink>ledyouhere, you may wish to change the link to point directly to the intended article.''</it>
	</col>
</row>
</table>
</p>
</bdy>
</article>
```

Figure II.16: Example of INEX 2009 Wikipedia XML document (file "8000.xml").

II.8.2. Topics

Topics are created in collaboration by the various INEX participants and they represent user needs in information. In INEX, two main categories of queries exist:

- The CO (Content Only) queries: they are queries in natural language, similar to those used in TREC. The keywords of the query can be grouped in the form of expressions and preceded by the operators '+' (meaning that the term is obligatory) or ' -' (meaning that the term should not appear in the elements returned to the user).
- The CAS (Content And Structure) queries: these queries make it possible to express constraints on the structure of the XML documents.

Each INEX topic (CO or CAS) is characterised by a whole of fields, which can be summarised in: the *Title* field which gives a formal definition of the query, the *Keywords* field which contains a set of keywords, and the *Description*and *Narration*fields, clarified in natural language, indicate the intentions of author (Sigurbjörnsson et al., 2005). By creating INEX topics, a certain number of factors should be taken into account. Thus, the topics would have:

- to be written by an expert (or somebody of familiar with) in the fields covered by the collection;
- to reflect real needs of the operational systems;
- to be varied;
- to be different in their cover;
- to be evaluated by the author of the subject.

The language used for the CAS queries expression (NEXI language) is an alternative of XPath (Clark & DeRose, 1999) and is detailed in (Trotman & Sigurbjörnsson, 2005). The formulation of the query is closely related to the associated retrieval task. Some examples of CAS queries are given in the following section.

II.8.3. Tracks

The INEX campaign includes several tasks. These tasks include the ad hoc track, which is regarded as a simulation of the use of a digital library. This library contains a set of documents on which the retrieval systems are running user queries. Queries (or topics) are of multiple types, the ad hoc task contains many sub tasks: Content Only, Content Only + Structure (CO+S) and

Content and Structure in its various forms (strict and vague): SSCAS, VSCAS, SVCAS and VVCAS. In the following sections, we will describe in detail each of these tracks. In addition to the ad hoc track, we find the interactive track, the NLP track (natural language processing), heterogeneous track, multimedia track, the Document mining track, and relevance feedback track.

II.8.3.1. Ad hoc Track

CO sub-task (*Content Only Task*). The purpose of this task is to respond to user queries of type CO by elements/XML documents. No indication of the granularity of the response to return is expressed, the CO application only contains keywords. The following table shows an example of INEX CO query:

```
<inex_topic topic_id="98" query_type="CO">
        <title> "Information Exchange" +"XML" "Information Integration"</title>
        <description>How to use XML to solve the information exchange
        (information integration) problem, especially in heterogeneous data
        sources ?</description>
        <narrative>Relevant documents/components must talk about techniques of
        using XML to solve information exchange (information integration) among
        heterogeneous data sources where the structures of participating data
        sources are different although they might use the same
        ontologies about the same content.
        </narrative>
        <keywords>Information exchange, XML, information integration,
        heterogeneous data sources</keywords>
</inex_topic>
```

TableII.2:Example of an INEX CO query

CO + S sub-task (Content Only + Structure). This type of queries attempts to simulate a user who does not know (or do not want to use) the actual structure of XML documents. This profile is likely to suit most users searching in collections of XML documents. Upon discovering that his CO query returned many irrelevant items, the user may decide to add structural constraints. This is similar to a user adding + and - operators to a Web search query when too many irrelevant pages are returned. At INEX, added these structural constraints (+S) are indicated using the formal syntax of NEXI (Trotman & Sigurbjörnsson, 2005). The following table shows a CO+S query of INEX 2005 campaign.

CAS sub-task (Content And Structure). A CAS query contains two types of structural constraints: where to look (support elements) and what elements to return (target elements). In earlier workshops INEX constraints on target elements have been interpreted strictly or vaguely, so that the support elements have always been vaguely interpreted (Sigurbjörnsson et al., 2005). These workshops led to create a discussion on how to interpret the support elements. There is the database view: all the structural constraints must be followed strictly (exact match), and the IR view: an element considered relevant if it satisfies the need for information irrespective of structural constraints.

Starting with INEX 2006, the CO+S queries was removed from the INEX evaluation campaign and only CO and CAS were adopted. Other ad hoc subtasks were defined: Thorough, Focused, Relevant in Context and Best in Context tasks.

The Thorough task asks systems to estimate the relevance of elements in the collection. The Focused task asks systems to return a ranked list of elements to the user. The Relevant in

Context task asks systems to return relevant elements clustered per article to the user. The Best in Context task asks systems to return articles with one best entry point (BEP) to the user.

```
<inex_topic query_type="CO+S">
        <title>formal logic reason UML diagrams</title>
        <castitle>//article[about(.//bb, Rumbaugh Jacobson Booch) and
        about(.//abs, formal methods)]//sec[about(.,formal logic reason UML
        diagrams)]</castitle>
        <description>I want to know about the application of formal methods and
        logics to reason about UML diagrams. Relevant items probably cite
        Rumbaugh, Jacobson, or Booch.</description>
        <narrative>My main interest is the application of formal methods and
        logics in software development. I choose to search for its application to
        UML diagrams because I think it is an interesting application area. To be
        relevant, a document/component must discuss the use of formal logics, such
        as first-order-, temporal-, or descriptionlogics, to model or reason about
        UML diagrams. I'm only interested in proper formal logics, Business-logics
        and Client-logics do not have a proof system and are therefore not
        considered to be formal logics. I think that sections are the most
        appropriate unit of retrieval for this fairly specific topic, since I'm
        not really interested in reading a lot about UML stuff in general. I want
        to focus in on the document parts that talk about logic. I think it is
        useful for the search engine to look for citation to the UML trio:
        Rumbaugh, Jacobson and Booch. Similarly think that it might be usefu to
        put the formal methods constraints on the abstract to stress that I'm only
        interested in this particular subset of UML articles. Of course a relevant
        article need not have this sort of reference or abstract, therefore the
        relevance of an element will be judged on basis of how well it explains
        the use of formal logics to model or reason about UML diagrams.
        </narrative>
</inex_topic>
```

TableII.3: Example of an INEX 2005 CO+S query

```
<inex_topic query_type="CAS">
        <castitle>//article[about(.//au,"Jiawei Han")]//abs[about(.,"data
        mining")]</castitle>
        <description>a synopsis of data mining papers by Jiawei Han</description>
        <narrative>I'm writinga short article about the impact of Jiawei Han on
        the field of data mining. Therefore I'm interested in finding a short and
        concise overview of his papers. I believe this is to be found in the
        abstracts of his papers. To be relevant, the component has to be the
        abstract, written by Jiawei Han, about "data mining". Any topics of data
        mining (e.g. association rules, data cube etc.) shouldbe considered as
        relevant.</narrative>
</inex_topic>
```

TableII.4: Example of an INEX 2005 CAS query

In our propositions and experiments (see chapter V) we use retrieval results returned for the Focused task. This task asks retrieval systems to find the most focused elements that satisfy an information need, without returning "overlapping" elements.

II.8.3.2. Relevance Feedback Track

The goal of this track is to study the query reformulation in the context of the IR in XML documents. The query reformulation should ideally consider not only the contents but also the structural constraints of the XML documents.

INEX uses the technique of residual collection to evaluate the effectiveness of relevance feedback approaches (Crouch, 2005). In this technique, all the examined XML elementsin the query reformulation process must be removedfrom the collection before the evaluation takes place. Under INEX directives, this means that not only each used or observed XML element in the query reformulation process but also all the descendants of this element must be removed from the collection. Whereas, all the antecedents of this XML element are maintained in the residual collection.

II.8.3.3. Other Tracks

Natural Language Processing Track. This track simulates a user asking the question to the information retrieval system in natural language. It examines the ability of a retrieval system to satisfy the need for information expressed in natural language.

Heterogeneous Track. This task is divided into two sub-tasks "HET.CO" and "HET.CAS". The purpose of the first subtask is to search for items in various collections based on their contents. While the second consists at using the structure or the implicit structure in addition to the content. This task has several collections of different DTDs (Sigurbjörnsson et al., 2005), and tries to answer the following questions:

- For CO queries, what are the feasible methods to determine the elements that can be considered as good answers?
- What are the methods that can be used to transform the structural constraints to other DTDs?
- The transformation must focus on the element names only or the content too?
- What are the evaluation criteria for heterogeneous collections?

Multimedia Track. The main purpose of the INEX multimedia task is to provide an evaluation platform for IRS that do not include only text in the retrieval process. Many structured document collections also contain other types of media, such as images, speech, and video. For Instance, INEX 2005 Multimedia task was based on "*The Lonely Planet WorldGuide*" which consists of 462 XML documents (Van Zwol et al., 2006).

II.8.4. Relevance Assessments

During runs, participants produce a list (or set) of XML elements for each of the topics. Tops 1500 retrieved elements are sent to the INEX campaign. Bids will be then distributed to the participants (as much as possible to the authors of topic) for assessments. Noting that the judgments regarding a topic must be performed by a single person (e.g. the author of the topic) (Lalmas & Piwowarski).

Relevance is defined in INEX by two dimensions(Lalmas & Piwowarski): Exhaustiveness: which describes the extent to which an XML element discusses the subject of the topic; and Specificity which describes how far the XML element focuses on the subject of the topic.

Several assessment tools have been made available to the various participants, for example, *XRAI* (INEX 2005)(Lalmas & Piwowarski), *EVALJ* (INEX 2007)(Lalmas & Piwowarski, 2007), *INEX_EVAL* (INEX 2009, 2010)(Geva et al., 2010; Lalmas & Piwowarski).

From INEX 2006 assessment process is done as follows: Assessors highlight text fragments containingrelevant information. They should read all article parts before deciding which text to highlight. If any part of an article is highlighted, this article is considered as relevant, which means that assessors should then select a "best entry point" (BEP) of thisarticle. For the CO+S topics, titles may contain structural constraints (using XPath expressions). These structural conditions should be ignored during assessment process. This means that assessors should consider the elements returned if they satisfy expressed information need with respect to the content criterion only.

II.8.5. Evaluation

To evaluate the performance of different developed retrieval systems in structured documents the INEX campaign adopted methods based on recall and precision measurements. It was inspired in large part of the work on the evaluation of IRS developed in the cranfield experiments and later in TREC (Hiemstra & Mihajlovic, 2005).

II.8.6. Relevance Assessments and Evaluation Metrics

In this section we give an overview on evaluation measures used from INEX 2002 to INEX 2010 campaign. We have already mentioned that the relevance in INEX is defined by two dimensions: exhaustivity and specificity. Most measurements use aggregate functions to produce the final result of the evaluation (Equation II.6).

$$F_{quant}(e,s) : ES \rightarrow [0,1] \tag{II.6}$$

Where ES represents all possible values for the pairs (e, s): ES = {(0,0), (1,1), (1,2), (1,3), (2,1) , (2,2), (2,3), (3,1), (3,2), (3,3)}

Each element can be marginally (1) enough (2) or very (3) exhaustive or specific, or irrelevant (denoted by the pair (0,0)).

II.8.6.1. INEX 2002 Measures: inex_eval

The measure of INEX 2002 (also called inex_eval) calculates the measure "precall" proposed by Kakade and Raghavan (Gövert & Kazai, 2002), using the probability that an XML element seen by the user is considered as relevant (P(rel/Retr)) (equation II.7):

$$P(rel/retr)(x) = \frac{x.n}{(x.n + esl_{x.n})} \tag{II.7}$$

Where $esl_{x.n}$ represents the length of supposed research (*expected search length*), i.e. the number of assumed non-relevant items returned until a recall point x is reached (Gövert & Kazai, 2002).

The assumed retrieval length is specified by the following formula:

$$esl_{x.n} = j + \left(\frac{s.i}{r+1} \right) \tag{II.8}$$

Where j is the total number of irrelevant elements in all levels preceding the final level; s is the number of relevant elements required in the final level to satisfy the recall point; i represents the number of irrelevant elements in the final level; and r represents the number of relevant elements in the final level (Hiemstra & Mihajlovic, 2005). The implementation of this measure requires the aggregation of the two relevant dimensions (E & S) to get a single value. Two types of functions have been used:

A "strict" aggregation for evaluating if a system is able to retrieve very specific elements:

$$f_{strict}(s,e) = \begin{cases} 1 & if \quad e=3 \quad and \quad s=3 \\ 0 & else \end{cases} \tag{II.9}$$

A "generalized" aggregation to evaluate the elements according to their degree of relevance.

$$f_{gen}(s,e) = \begin{cases} 1 & if \quad (e,s) = (3,3) \\ 0.75 & if \quad (e,s) \in \{(2,3),(3,\{2,1\})\} \\ 0.5 & if \quad (e,s) \in \{(1,3),(2,\{2,1\})\} \\ 0.25 & if \quad (e,s) \in \{(1,2),(1,1)\} \\ 0 & if \quad (e,s) = (0,0) \end{cases} \tag{II.10}$$

II.8.6.2. INEX 2003 Measures: inex_eval_ng

The INEX 2003 measures (also called inex_eval_ng) try to overcome the problem caused by the 2002 measure which is the nesting of elements in the retrieval results by incorporating the size of items and nesting in the definition of recall and precision. However it does not address the problem of nesting of the XML elements in the evaluation. Nesting problem is surpassed by considering only incrementing the size of already seen elements. This measure implies that the relevant information is uniformly distributed within an element, which has not been proven in practice (Hiemstra & Mihajlovic, 2005).

Recall and precision formulas for "inex_eval_ng" measures are:

$$recall_0 = \frac{\sum_{i=1}^{k} e(c_i).\frac{|c_i'|}{|c_i|}}{\sum_{i=1}^{N} e(c_i)} \tag{II.11}$$

$$precision_0 = \frac{\sum_{i=1}^{k} s(c_i).\left|c_i^{'}\right|}{\sum_{i=1}^{N}\left|c_i^{'}\right|} \quad \text{(II.12)}$$

where $\mathbf{c_1, c_2, ...c_n}$ represent an ordered list of results; **N** is the total number of items in the collection; $\mathbf{e(c_i)}$ and $\mathbf{s(c_i)}$ denote respectively the values of exhaustivity and specificity attributed to $\mathbf{c_i}$; $\mathbf{|c_i|}$ represents the size of $\mathbf{c_i}$; and $\mathbf{|c_i'|}$ is the size of the item previously viewed by the user which can be calculated as follows:

$$\left|c_i^{'}\right| = \left| c_i - \bigcup_{c \in C[1,n-1]} (c) \right| \quad \text{(II.13)}$$

Where **n** is the rank of the element **ci** and **c[1,n-1]** is the set of elements to return between positions **[1,n-1]**. Aggregation functions are defined so that they provide a separate processing exhaustivity and specificity:

$$f_{quant}^{'}(e) : E \to [0,1] \quad and \quad f_{quant}^{'}(s) : S \to [0,1] \quad \text{(II.14)}$$

In the case of strict aggregation, the only function takes as values e = 3 and s = 3, respectively. For the case of the generalized aggregation functions are defined as follows:

$$f'_{gen}(e) = e/3 \quad \text{(II.15)}$$

And

$$f'_{gen}(s) = s/3 \quad \text{(II.16)}$$

The problem of this measurement is the separation of the two dimensions of relevance (Hiemstra & Mihajlovic, 2005).

II.8.6.3. INEX 2004 Measures: Specificity and Exhaustivity Oriented Functions

Based on the discussion conducted during INEX 2003 about aggregate functions and disadvantages of measures of this campaign, two classes of aggregate functions were introduced for INEX 2004 (defined).

The exhaustivity oriented functions applied to the strict aggregation:

$$f_{e3\_s321}(s,e)=\begin{cases}1 & if \quad s\in\{3,2,1\} \ \ and \ \ e=3\\ 0 & else\end{cases} \tag{II.17}$$

$$f_{e3\_s32}(s,e)=\begin{cases}1 & if \quad s\in\{3,2\} \ \ and \ \ e=3\\ 0 & else\end{cases} \tag{II.18}$$

And similarly, specificity oriented functions applied to the strict aggregation:

$$f_{s3\_e321}(s,e)=\begin{cases}1 & if \quad e\in\{3,2,1\} \quad and \quad s=3\\ 0 & else\end{cases} \tag{II.19}$$

$$f_{s3\_e32}(s,e)=\begin{cases}1 & if \quad e\in\{3,2\} \ \ and \ \ s=3\\ 0 & else\end{cases} \tag{II.20}$$

However, this measure suffers from the problem of nesting elements (Hiemstra & Mihajlovic, 2005).

II.8.6.4. INEX 2005 Measures: XCG (eXtended Cumulated Gain)

The XCG measures (eXtended Cumulative Gain) are extensions of the cumulative gain (CG) proposed by Jarvelin Kekäläinen (Järvelin & Kekäläinen, 2002). This cumulative gain can be calculated at each position ***i*** according to the following formula:

$$CG[i]=\sum_{j=1}^{i} G[j] \tag{II.21}$$

In the above formula, the cumulative gain at position ***i*** is the sum of relevance scores of its lower positions (Hiemstra & Mihajlovic, 2005). The function that sets the relevancy score for an item using the XCG measure is:

$$rv(c_i) = f(quant(assess(c_i))) \quad \text{(II.22)}$$

Where assess($\mathbf{c_i}$) is the function allowing to return the judgment pair of the element $\mathbf{c_i}$, and quant(assess($\mathbf{c_i}$)) represents the aggregate function.

The function ***f*** has three variants:

- in the case where the element $\mathbf{c_i}$is not yet considered f(x)=x=quant(assess(c_i));
- in the case where the element $\mathbf{c_i}$ is seen: f (x) = (1-α) * x; where α is a factor which simulates the behaviour of the user in accordance with the already seen elements.
- Finally, in the case where part of $\mathbf{c_i}$ is already seen:

$$f(x) = \alpha . \frac{\sum_{j=1}^{m} (rv(c_j) . |c_j|)}{|c_j|} + (1-\alpha) * x \quad \text{(II.23)}$$

Where ***m*** is the number of child nodes of the node $\mathbf{c_i}$considered as relevant (Hiemstra & Mihajlovic, 2005).

II.8.6.5. INEX 2006 Measures: nxCG (normalised eXtended Cumulated Gain)

The INEX 2006 ad hoc track covers four retrieval tasks: focused, thorough, relevant in context, and best in context.

Various sets of measures are used to evaluate these different tasks: XCG measures for the thorough and focused tasks; generalized precision measure for the context retrieval task; BEPD and EPRUM measure for the best in context.

Relevance assessments of INEX 2006 are obtained by assessors highlighting relevant text fragments in the documents. XML elements that contained some highlighted text were then considered as relevant (to varying degree). For each relevant XML element, the size of the contained highlighted text fragment is recorded as well as the total size of the element (in number of characters). These two statistics form the basis of calculating an XML element's relevance score, which in INEX 2006 corresponds to its specificity score(Lalmas et al., 2007).

The specificity score, $spec(e_i) \in [0 , 1]$, of an element ei is calculated as the ratio of the number of highlighted characters contained within the XML element, $rsize(e_i)$, to the total number of characters contained by the element, $size(e_i)$:

$$spec(e_i) = \frac{rsize(e_i)}{size(e_i)} \quad \text{(II.24)}$$

The normalized cumulated gain nxCG[RCV] measure (XCG family of measures), was used in the evaluation of the focused task. System performance is reported at several rank cutoff values (RCV). For a given topic, the normalized cumulated gain measure is obtained by dividing a retrieval run's xCG vector by the corresponding ideal xCI vector (Lalmas et al., 2007):

$$nxCG[i] := \frac{xCG[i]}{xCI[i]} \tag{II.25}$$

xCG[i] takes its values from the full recall-base of the given topic with $i \in [0, 1500]$ where 1500 is the maximum length of a result list that participants could submit. *xCI[i]* takes its values from the ideal recall-base and *i* ranges from 0 and the number of relevant XML elements for the given topic in the ideal recall-base. The gain values *xI[j]* used in *xCI[i]*formula are given by the following equation.

$$xI[j] = spec(e_j) \tag{II.26}$$

The gain values used in *xCG[i]*are normalized as follows. For the *j-th* retrieved element, where *j* ranges from 1 to *i*:

$$xG_{norm}[j] = \min(xG[j], xG[j_{ideal}] - \sum_{s} xG[k]) \tag{II.27}$$

Where:

- $xG[\cdot]$ is given by the following equation:

$$xG[i] = spec(e_i) \tag{II.28}$$

- j_{ideal} is the rank of the ideal element that is on the same relevant path as the *j-th* relevant element, and
- *S* is the set of elements that overlap with that ideal element and that have been retrieved before rank *j*.

For a given rank *i*, *nxCG[i]* reflects the relative gain the user accumulated up to that rank. nxCG calculated by taking measurements on both the system and the ideal rankings' cumulated gain curves along the vertical line drawn at rank *i*. Here, rank positionis used as the control variable and cumulated gain as the dependent variable(Lalmas et al., 2007).

II.8.6.6. INEX 2007, 2008 and 2009 Measures: Interpolated Precision (iP) & MAgP

INEX 2007 measures are based on the amount of highlighted text retrieved, leading to natural extensions of the well-established measures of precision and recall. The Focused Task is

evaluated by using interpolated precision at 1% recall (*iP[0.01]*) in terms of the highlighted text retrieved.

Let p_r be the document part assigned to rank *r* in the ranked list of document parts *Lq* returned by a retrieval system for a topic *q* (at INEX 2007, Lq| = 1500 elements or passages). Let *rsize*(p_r) be the length of highlighted (relevant) text contained by p_r in characters (if there is no highlighted text, rsize(p_r) = 0). Let size(p_r) be the total number of characters contained by p_r, and let *Trel(q)* be the total amount of (highlighted) relevant text for topic *q*.Trel(q) is calculated as the total number of highlighted characters across all documents, i.e., the sum of the lengths of the (non-overlapping) highlighted passages from all relevant documents(Kamps, Pehcevski, Kazai, Lalmas, & Robertson, 2008).

Precision at rank *r*is defined as the fraction of retrieved text that is relevant:

$$P[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{size(p_i)} \tag{II.29}$$

To achieve a high precision score at rank r, the document parts retrieved up to and including that rank need to contain as little non-relevant text as possible. Recall at rank *r*is defined as the fraction of relevant text that is retrieved:

$$R[r] = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)} \tag{II.30}$$

To achieve a high recall score at rank r, the document parts retrieved up to and including that rank need to contain as much relevant text as possible (Kamps et al., 2008).

An issue with the precision measure P[r] given in Equation II.29 is that it can be biased towards systems that return several shorter document parts rather thanreturning one longer part that contains them all (issue well known at passage retrieval task of TREC). Since the notion of ranks is relatively fluidfor passages, INEX 2007 opted to look at precision at recall levels rather than at ranks.Specifically, by using an interpolated precision measure *iP[x]*, which calculatesinterpolated precision scores at selected recall levels.

$$iP[x] = \begin{cases} \max\limits_{1 \le r \le |L_q|} \left(P[r] \wedge R[r] \ge x\right) & if\ x \le R\left[|L_q|\right] \\ 0 & if\ x > R\left[|L_q|\right] \end{cases} \tag{II.31}$$

II.9. Conclusion

In this chapter, we presented the various elements of the XML Information Retrieval. Initially, we gave an outline on the structured (XML) documents. Thus, we presented some XML IR challenges relating to the structural aspect of these documents. Different indexing approaches and models were proposed in the literature, as well as languages of interrogation of the corpora of XML documents. These languages make it possible to the users to express their needs through two types of queries: Content only queries and content and structurequeries. The second type of queries supposes that the user has an idea on the structure of the XML document, therefore it can indicate the type of the unit of information to be returned. Whereas in the first type, it is the task of the IR system to decide of the information granularity to return.

These last years, several IR systems were developed. Each one of these systems has its query processing method. These systems take part each year in an evaluation campaign called INEX.

Historically, Web information retrieval is known to be the first public used field exploiting the links in the information retrieval process.

The next chapter will be devoted to synthesize all of the concepts, models and methods proposed for the incorporating of the link evidence in the field of information retrieval on the Web.

Chapter III.

Links in Web Information Retrieval

III.1. Introduction

The major focus of IR research is on developing strategies for identifying "relevant" documents in respect to a given query. In traditional IR, evidence of relevance is typically mined from textual evidence contained in these documents. It is based on ranking documents according to their degree of relevance using measures like:term similarity or term occurrence probability. On the Web, however, other sources of evidence can be incorporated to retrieve relevant information in documents. For Instance, Web document metadata can be easily collected and used, such as statistics and documentproperties (e.g., size, date, etc.), which can be used to combined with the document term-based relevance. Hyperlinks, is considered as the most prominent source of evidence in the context of Web documents, which have been the subject of many studies exploring retrieval strategies based on link use.

In this chapter, we review the main concepts, methods and models proposed in the Web IR literature, in particular the use of link information as a source of evidence in the retrieval process. This chapter is organised as follows:

- SectionIII.2presents the concepts of Hypertext and the Web information retrieval field;
- SectionIII.3described the value of link evidence in the IR field;
- SectionIII.4 described an outline of the most well-known link-based ranking algorithms (mainly used in the Web IR), for instance, Pagerank, HITS and SALSA;
- SectionIII.5 presents some aspects related to links and search engines;
- SectionIII.6Will conclude this chapter with a brief discussion.

III.2. Hypertext and Web Information Retrieval

The creation of hyperlinks is considered as one of the great advantages of digital documents. It allow the readers to jump directly from one document to another, related document, without having to search for a physical copy of the referenced document. Ideas about a large information networks of interlinked documents date back as far as the 1930s, when *Paul Otlet* envisioned a new form of globally accessible encyclopaedia based on linked documents (Rayward, 1994).

Before the advent of the Web, there were many ideas about using hyperlinks for retrieval of hypertext media. Most of these approaches considered the topical relatedness of linked documents. In other words, they hoped to use links to determine the topical relevance of documents (Koolen, 2011). The proposed approach attempt to prove that hyperlink information (text and target) had something new to offer for IR experimentation.

Within a hypertext collection, the reader is expected to navigate through links to create their knowledge about a topic of interest. The presence of hyperlinks puts a strain on the assumption adopted for the Cranfield experiments that the relevance of a document is independent of other documents in the collection (Koolen, 2011). Links were seen as valuable evidence for identifying relevant documents, but they also introduced interesting problems for the IR community.

The Web offers a rich context of information which is expressed through the hyperlinks. A link from page *p* to page *q* denotes an endorsement for the quality of page *q*. The Web can be seen as a network of recommendations which contains information about the authoritativeness of the pages.

Using hyperlinks to enhance retrieval process is based on the notion that hyperlinks connect related documents (similar topics) and thus can provide additional information. Link-based retrieval strategies explore several methods to incorporate hyperlink information with document contents. The aim of the ranking model is to extract the relevant information and produce a ranking that reflects the relative authority of Web pages.

In the context of hypertext, where the document collection usually consists of homogeneous documents on a single topic that are linked together for the purpose of citation, external content introduced by hyperlinks tends to be of high quality and usefulness. On the Web, however, where hyperlinks connect documents of varying quality and content for various purposes, document enrichments via hyperlink can sometimes introduce noise and degrade the retrieval performance (Koolen, 2011).

In Web information retrieval, characterized by the size of available data and the special behaviour of users, the role of the ranking model becomes critical. In the current context, Web search engines can return results containing thousands or millions of pages for a given query. Many studies about the user behaviour on the retrieval results observed that most of Web users do not look beyond the first top pages of results. Thus, it is important for the ranking model to return the desired results within the top few pages.

The specific characteristics of the Web have made traditional information retrieval models less effective for Web IR. These characteristics, related to Web, Web pages and Web users, can be summarized in the following points:

- **Index size.** The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Currently, Google index far exceeds the 100 million gigabytes[12].
- **Dynamic Web.** The Web content keeps growing and changing everyday while traditional IR techniques were mainly designed for static text databases.
- **Heterogeneity.** Web documents are inherently different: plain text, pdf, word documents, pictures, videos, audios, flash animations, interactive documents, etc.

[12] http://www.google.com/insidesearch/howsearchworks/crawling-indexing.html

- **Dynamic content.** As more and more Websites are using server side scripting to dynamically generate content, even the same URL contains different content when accessed at different time.
- **Multilingual.** More than 100 languages are used on the Web and even one page could use two or more different languages.
- **Notable link structure.** The presence of hyperlinks is one of the fundamental differences between Web documents and classical text databases. This link structure could be used to infer the general importance of a Web page.
- **Redundancy.** We can find on the Internet, tens or thousands of pages sharing the same content
- **User variance.** Web users have their own preferences and information needs. A good search engine must take into account user profile to return the best retrieval results.
- **User behaviour.** It is estimated that nearly 85% users only look at the first screen of the retrieval results. Thus, only top ranked results are meaningful in this context for most of the Web users.

Web IR, goes beyond the textual content of the document corpus and leverages numerous sources of evidence such as link structure, usage patterns, etc. While traditional IR depends solely on the textual content of documents, Web IR goes beyond the textual content and utilizes other sources of evidence.

III.3. The Value of Links for Information Retrieval

III.3.1. Web Mining

A web page may be more or less relevant to a query depending on its link structure features, including its indegree and outdegree, in addition to being on its content, which is the case of textual resources. The main arguments are: first, the Web pages are connected by hyperlinks per se; second, Web users do necessarily navigate between pages through the link structure, which means that the retrieval process is influenced by link evidence; third, assessments about the relevance of a Web page strongly depend on the assessments given on the pages previously seen. Thus the analysis of the Web link structure, considered as one kind of Web mining, plays a very important role in Web IR.

Web mining is the application of data mining and other information techniques to discover patterns and useful knowledge from the Web. The learned knowledge can be used to improve the efficiency and effectiveness of Web users' accessing the Web. According to analysis targets, web mining can be divided into three different types, which are: Web usage mining, Web content mining and Web structure mining[13].

Web structure mining process of discovering structure information from the Web. According to the type of web structural data, web structure mining can be divided into two brands: (a) Extracting patterns from hyperlinks in the web; (b) Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage. The

[13] http://en.wikipedia.org/wiki/Web_mining

structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.

When comparing web mining with traditional data mining techniques, there are three main differences to consider[14]:

- *Scale*: In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.
- *Access*: When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights.
- *Structure*: A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

Web Mining can contribute greatly to search technologies by designing new ranking algorithms based on discovered knowledge about the Web content, structure and users (Koolen, 2011).

III.3.2. Link Analysis

The analysis of the hyperlink structure of the web has led to significant improvements in web information retrieval. Severalweb rankingalgorithms exploit the linkage information inherent in the structure of the web in the retrieval process. The most prominent algorithmsusing this source of evidenceare:PageRank, HITS and SALSA. Also, there are numerous alternative link-based algorithms for ranking web results which represent in general improvements or variations of: PageRank, HITS and SALSA, for instance, BHITS, PHITS and TrustRank. These link-based ranking algorithms belong to two classes: query-dependent link-based ranking methods and a query-independent link-based ranking methods.

The link based approaches consider the web as a directed graph $G(V, E)$, where V represents the set of nodes and E the set of edges. Each Web page is represented as a node of the link graph and each hyperlink between two Web pages is modelled by directed edge.

The task of generating authoritative ranking of Web documents has become the most successful application of link analysis. It comprises many assumptions:

- The first assumption is that hyperlinks embed information ofuser's assessments about Web pages, which opens a rich potential for link analysis and Web IR field. Other than statistical features of links (number of incoming and outgoing links), the measure of general importance of Web pages computed by considering links is useful for Web IR since it contributes to the notion of relevance.

[14] http://www.scaleunlimited.com/about/web-mining/

- The second assumption is related to the semantic aspect of pages (relatedness), i.e. If a page *B* and *A* are connected by a link, then the probability that they are on the same topic is higher than if they are not connected(Monika Rauch Henzinger, 2000).

Link Analysis is used in many aspects of Web retrieval. For Instance: in (K. Bharat, Broder, Dean, & Henzinger, 2000), link structure analysis is used in the task ofmirrored Web hosts detection, i.e. mirror hostsare of very similar link structure; In (K. Bharat, Chang, Henzinger, & Ruhl, 2001), authors use the link graph of Web hosts to study the link dynamics of theseWeb hosts and domains. Similar techniques tryto estimate the coverage and measure the index quality of many search engines(Monika R Henzinger, Heydon, Mitzenmacher, & Najork, 1999).Link analysis is also used together with content analysis in automatic Web page categorization.

III.4. Link Based Ranking Methods

Link-based ranking algorithms use the link structure to determine the importance of nodes (elements or documents) in a link graph. The best-known algorithms are: degree-based (indegree/outdegree) and propagation-based algorithms (Pagerank, HITS, SALSA etc.) (Koolen, 2011).

Degree-based algorithms are the simplest way to derive information from the link structure. It consists at counting links incident to a page, called the link degree. Propagation based algorithms consist at propagating some kind of score from one node of the link graph to another via links. The best-known propagation algorithms are PageRank, HITS, SALSA, etc.

In this section we will describe the three Well Known algorithms used in the Web IR context, i.e., PageRank(Brin & Page, 1998), HITS (Kleinberg, 1999) and SALSA (Lempel & Moran, 2000). These algorithms are classified into two classes: ***(a)*** Global context methods which includes PageRank; and;***(b)*** Topic or local context methods which includes HITS, SALSA and a variant of PageRank (i.e. topic sensitive PageRank). The global context means that the computation of document scores is performed offline on the entire linked collection. The topic context (or "Query-dependent" context) means that the computation is done at query evaluation on a subset of retrieved documents for a given topic.

III.4.1. Precedents of Link Analysis

The advent of link-based approaches predates the Web and can be traced back to citation analysis in the field of bibliometrics and to hypertext research. The idea of using citation information to find documents related to each other was investigated well before the Web(Koolen, 2011) and has been the subject of much work in Sociometrics community. Kessler (Kessler, 1963) has introduced a method for grouping scientific literature based on bibliographic coupling units.

Two measures of document similarity based on citations were proposed in bibliometrics[15]: bibliographic coupling (or co-reference)(Kessler, 1963) and co-citation(Small, 1973). Bibliographic coupling represents the number of documents cited by both document *p* and *q*. Co-

[15]Bibliometrics is the study of written documents and their citation structure

citation represents the number of documents that cite both *p*&*q*. Shaw (1991a, 1991b) deploy these measures by using a combination of text similarity, bibliographic coupling, and co-citation as part of a graph-based clustering algorithm to improve retrieval accuracy. These two notions are also adapted and used in Web link analysis.

Ding et al. (Ding, He, Husbands, Zha, & Simon, 2002)have noted the underlying connection between HITS algorithm and two bibliometrics concepts: co-citation and co-reference. They observed that in information retrieval field, co-citation occurs when two nodes (pages in the context of Web IR) share a common inlinking node, while co-reference means that two nodes share a common outlinking node. Authors (Ding et al., 2002)showed that the authority matrix L^TL of HITS algorithm has a direct relationship to the concept of co-citation, while the hub matrix LL^T is related to co-reference.

III.4.2. INDEGREE

One of the simplest ways to derive information from the link structure is to count links incident to a page, called the link degree(Koolen, 2011). Many statistical features can be used, for instance: the incoming link degree (or in-degree) which represents the number of incoming to a given graph node; the outgoing link degree (or out-degree) which represents the number of outgoing links from a given graph node; or the combination of these two degrees.

The INDEGREE Algorithm is a simple heuristic that can be considered as the predecessor of all link analysis ranking algorithms. Its principal is to rank the pages according to their popularity (or visibility) (Marchiori, 1997). This visibility is measured by the number of incoming links (inlinks) of a page. This algorithm is referred as the INDEGREE algorithm, since it ranks pages according to their in-degree in the graph of links. For every node i:

$$\text{Indegree}(a_i) = |\text{Inlinks}(a_i)| \qquad \text{(III.1)}$$

This simple heuristic was applied by several search engines in the early days of Web IR (Marchiori, 1997). In his proposition of HITS, Kleinberg (Kleinberg, 1999) makes a convincing argument that the INDEGREE algorithm is not sophisticated enough to capture the authoritativeness of a node, even when restricted to a query dependent subset of the Web.

III.4.3. HITS

Hyperlink-Induced Topic Search (HITS), also known as hubs and authorities, is a link analysis algorithm that rates Web pages, proposed and implemented by Jon Kleinberg (Kleinberg, 1999) in the search engine of IBM. Kleinberg distinguished between two types of Web pages which pertain to a certain topic: authorities (good sources of content) and hubs (good sources of links). Hub pages are pages that "pull together" authorities on a given topic, and allow to throw out unrelated pages of large indegree (universally popular pages like Yahoo!). Hubs and Authorities exhibit a *mutually reinforcing relationship*. A good hub page is one that points to many good authorities; a good authority page is one that is pointed to by many good hub pages. Consequently we appear to have a circular definition of hubs and authorities; that will turn this into an iterative computation. Thus, the first line of the following equation sets the hub score of page *v* to the sum of the authority scores of the pages it links to, and the second line sets the authority score of the same page *v* to the sum of the hub scores of the pages linking to it.

$$h(v) = \sum_{v \to y} a(y)$$
$$a(v) = \sum_{y \to v} h(y) \qquad \text{(III.2)}$$

These two formulas can be interpreted as follows, if *v* links to pages with high authority scores, its hub score increases. Inversely, if page *v* is linked to by good hubs, its authority scores increases.

Hub and authority scores are computed for a subset of web pages selected as follows:

- Extract a root set of pages by applying a term-based search engine,
- From this root set we derive a base set which consists of:
  - (a) Pages in the root set, generally the top *N* pages for a given topic;
  - (b) Pages which point to a page in root set;
  - (c) Pages which are pointed to by a page in the root set.

These dual rankings of the HITS algorithm are one of its advantages for IR. HITS presents two ranked lists to the user: one with the most authoritative documents related to the query and the other with the most "hubby" documents. A user may be more interested in one ranked list than another depending on the retrieval task. HITS designed the overall Web information retrieval problem as a small problem, by finding the dominant eigenvectors of small matrices such as the size of these matrices is very small relative to the total number of Web documents(Langville & Meyer, 2005).

However, there are some clear drawbacksof the HITS link analysis algorithm. The most known problem with HITS is its susceptibility to spamming. In other words, a user can slightly influence the authority and hub scores of his page by adding links to and from his webpage. For a webpage owner, adding outgoing links from a page is much easier than adding incoming links to that page. Therefore, influencing one's hub score is not difficult. Since hub scores and authority scores are computed inter-dependently, an authority score will increase as a hub score increases. Another problem with HITS is the problem of topic drift. In the "building the neighbourhood graph"step for a query it is possible that a very authoritative yet off-topic document be linked to a document containing the query terms. This very authoritative document can carry so much weight that it and its neighbouring documents dominate the relevant ranked list returned to the user, skewing the results towards off-topic documents. Henzinger and Bharat (K. Bharat & Henzinger, 1998; K. A. Bharat & Henzinger, 2000)proposed a solution to the problem of topic drift, by weighting the authority and hub scores of the nodes in theneighbourhood graph by a measure of relevancy to the query.

TKC Effect

The TKC (Tightly Knit Community) effect is probably one of the main reasons that the algorithm HITS has not been as successful as the PageRank algorithm. Like the latter, the HITS algorithm was originally proposed in information retrieval for classification ("ranking") of documents on the web. In its "full" version, HITS (Kleinberg, 1999) built initially a graph of documents from a user query and, secondly, computes authority and hub scores. Several

researchers (Lempel & Moran, 2000)that studied the HITS algorithm have noticed that documents classified as important do not typically address only one topic of the user query. Worse, it has even been observed that in some cases, HITS ranks as important documents that are not relevant to the original query (a problem known as the topic drift (K. Bharat & Henzinger, 1998).

A tightly-knit community is a small but highly interconnected set of Pages. The TKC effect occurs when such a community scores high in link ranking algorithms, even though the pages in the TKC are not authoritative on the topic, or are relevant to just one part of the query topic. (Lempel & Moran, 2000) indicates that the Mutual Reinforcement approach (i.e. HITS) is vulnerable to this effect by rankingsometimes pages of a TKC set in unjustified high positions.

In (Lempel & Moran, 2000),authors' experiments confirm thatthe Mutual Reinforcement approach ranks highly authorities on only one aspect of the query.

III.4.4. PageRank

PageRank algorithm was proposed by S. Brin and L. Page (Brin & Page, 1998). It applied the citation analysis principal on the graph of the web. Quoting from the original Google paper, PageRank algorithm is defined as follows:

We assume page A has pages $T_1...T_n$ which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n))$$

The PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one. *PR(A)* can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

The main idea of the PageRankalgorithm is to simulate the behaviourof a user navigating randomly on the Web. The probability to visit a given page is even higher if this page is pointed by many other pages. Considering that page gives some authority to another page by linking to it, the probability of navigating of the random surfer on a page indicates the degree of it relevance. The computation of PageRank is doing following this formula:

$$PageRank(p) = \frac{1-d}{docs\_nbr} + d * \sum_{(q,p)\in links} \frac{PageRank(q)}{OutLinks(q)} \qquad \text{(III.3)}$$

Where:

- *d* represent the damping factor (usually fixed at 0.85 by Google);
- *docs_nbr* represent the number of pages on the Web collection;
- *links* represent all the pairs *(i, j)* of links within the Web collection such that the page *i* contains a link to the page *j*.
- *OutLinks(q)* represent the number of outgoing hyperlinks on page *q*.

The random surfer visits a web page with a certain probability which derives from the page's PageRank. The probability that the random surfer clicks on one link is solely given by the

number of links on that page. The probability for the random surfer reaching one page is the sum of probabilities for the random surfer following links to this page. This probability is reduced by the damping factor *d*.

The major well known problem in PageRank algorithm is: the topic drift problem, due to the importance of determining an accurate relevancy score(Langville & Meyer, 2005). Bharat et al. (K. Bharat & Mihaila, 2001)succinctly state this weakness of PageRank, by: "*Since PageRank is query-independent, it cannot by itself distinguish between pages that are authoritative in general and pages that are authoritative on the query topic.*"

Regarding the strengths of PageRank, we can mention its virtual imperviousness to spamming(Langville & Meyer, 2005). It is very hard for a webpage owner to add incoming links into his page from other important pages (see section III.4.3). Some studies (Chien, Dwork, Kumar, & Sivakumar, 2001)have demonstrated that if the owner succeeds to add incoming links into his page from other important pages, its PageRank will increase. However, this increase will likely be insignificant since PageRank is a global measure. However, authority and hub scores of HITS' algorithm are derived from the"local neighbourhood graph" and small increases in the number of incoming links or outgoing links will have a greater relative impact.

III.4.5. SALSA

SALSA (Stochastic Approach for Link Structure Analysis), was proposed by Lempel and Moran(Lempel & Moran, 2000). Theirproposition is similar to the HITS algorithm described above (i.e. they employ the same meta-algorithm): it attempts to determine the best pages for a given topic by characterizing them as hubs and authorities. SALSAis based upon the theory of Markov chains, and relies on the stochastic properties of random walks performed on collection of Web pages.

Authors combine the theory of random walks(Brin & Page, 1998) with the notion of hubs and authorities, and analyse two different Markov chains: a chain of hubs and a chain of authorities. A bipartite undirected graph is constructed from a given collection, one side of the bipartite graph represents the hubs and the other side represents the authorities.Its implementation follows the HITS approach of identifying topic-driven neighbourhood graphs while replacing the iterative algorithm of mutual reinforcement approach with a non-iterative stochastic approach to identify hubs and authorities.

Authors define two stochastic matrices, which are the transition matrices of the two Markov chains:

(1) The hub matrix, defined as follows:

$$\tilde{h}_{i,j} = \sum_{k:k\in S(i)\cap S(j)} \frac{1}{|S(i)|}\cdot\frac{1}{|E(k)|} \qquad \text{(III.4)}$$

(2) The authority matrix, defined as follows:

$$\tilde{a}_{i,j} = \sum_{k:k\in E(i)\cap E(j)} \frac{1}{|E(i)|}\cdot\frac{1}{|S(k)|} \qquad \text{(III.5)}$$

Where:

- *E(i)* describes all pages in the graph pointing to the page *i*;
- *S(i)*describes all pages we can achieve from the page *i*.

A transition probability $\tilde{a}_{i,j} > 0$ implies that a certain page *k* points to both pages *i* and *j,*and hence page *j* can be reached from the page *i* in two steps: by browsing the link *k* to *i* in the opposite direction and then following the link from page *k* to page *j*.

As SALSA was developed by combining some of the best features of both HITS and PageRank, it has many strengths. Unlike HITS, SALSA is not victimized by the topic drift problem, related to the "TKC" problem (Langville & Meyer, 2005). Lempel and Moran (Lempel & Moran, 2000) propose an approach to resolve the TKC effect (i.e., diffusion effect, topic drift) of HITS. Also, SALSA algorithm is less susceptible to spamming problem since the coupling between hub and authority scores is much less strict, instead of HITS which is susceptible due to the interdependence of hub and authority scores.

The most important reason for the claimed effectiveness of SALSA lies probably in its careful filtering of links in the link graph formulation stage. Lempel and Moran, proposing that filtering out of "non-informative" links is one of the most crucial steps in link analysis. This step eliminate about 38% of links to arrive at a high-quality link graph by ignoring related-domain links. They suggest that link differentiation by link propagation is not as important when the link graph is of high quality (Yang, 2005).

III.4.6. Other Link Analysis Algorithms

The innovative work of Kleinberg (Kleinberg, 1999), and Brin and Page (Brin & Page, 1998) was followed by several extensions and modifications. Bharat and Henzinger (K. Bharat & Henzinger, 1998) and Chakrabarti et al.(Chakrabarti et al., 1998)consider improvements on the HITS algorithm by using textual information to weight the importance of nodes and links. (Rafiei & Mendelzon, 2000) consider a variation of the HITS algorithm that uses random jumps, similar to SALSA.The same algorithm was also considered by Ng et al. (Ng, Zheng, & Jordan, 2001) and called"Randomized HITS". Other Extensions of the HITS algorithm that use multiple eigenvectors were also proposed in (Ng et al., 2001).

III.5. Search Engines and Link Based Retrieval Algorithms

For several years, search engines have been used by web users since the Internet became part of daily life to search for relevant information contained in Web resources. The majority of current Web users found that Google provided good search results in response to their queries. For them Google delivered better results compared to traditional search engines.

One of the main differences between modern search engines and traditional ones is the adoption of link-based ranking algorithm in ordering Web documents. Traditional Web search engines often provide poor search results since the use only text-based in the ranking process. Google has claimed that it is its link-based ranking algorithm, i.e. Pagerank that has made the quality of its retrieval results superior.

III.6. Conclusion

This chapter describes the main concepts and algorithms related to the use of links in the context of Web information retrieval. Thus, we have explained the concept of hypertext and information retrieval on the Web.

We have emphasized the value of links in information retrieval through both fields: Web mining and Link analysis.

We have focused on well-known link analysis algorithms proposed, for instance, PageRank, HITS and SALSA.

Nowadays, new challenges of the IR field have appeared by the growing quantity of available structured information resources, principally collections of XML documents. Therefore, the logical structure of XML documents, representing a new source of evidence, has been exploited to retrieve XML elements at different levels of granularity. Instead of classical information retrieval approaches that focus on seeking unstructured content, XML information retrieval combines both textual and structural information to perform different IR tasks. A number of approaches taking advantage of the two types of information (textual and structural) have been proposed and are essentially based on traditional information retrieval models adapted to process the content part of the XML documents context.

Despite the popularity of links in the web and the conceptual proximity between HTML and XML links, only few approaches have exploited links connecting XML documents in XML IR context.

We will focus, in the next chapter, on the essentials of the work proposed to exploit the link evidence in the context of XML IR.

Chapter IV.

State of the Art on the Use of Links in XML Information Retrieval

IV.1. Introduction

In this chapter we present the main approaches for the use of link evidence in the XML information retrieval literature. Links have been widely exploited in the context of Web retrieval. Several algorithms were proposed, including PageRank(Brin & Page, 1998), HITS(Kleinberg, 1999) and SALSA(Lempel & Moran, 2001). Despite their popularity in web context, only few approaches, exploiting this source of evidence, have been conducted in the context of XML retrieval. These approaches can be classified into three categories: First, approaches whichanalyse the structure and nature of links in collections of XML documents. This category of approaches attempt to determine which characteristics of links (number, distance, anchors) can be exploited as a source of evidence in XML IR. Second, approaches proposing link detection strategies, mostly under the "Link-The-Wiki" task ofINEX campaign. These approaches focused on automatically linking orphan pages to already existing Wikipedia pages. Third, approaches exploiting links to re-rank the list of elements initially returned by an information retrieval system. In the context of our thesis we focused on the third category of approaches.

The present chapter is organized as follows: the first section is devoted to the description of the approaches studying the structure and the nature of INEX Wikipedia links (sectionIV.2). In sectionIV.3we review some link detection approaches. SectionIV.4discusses approaches using the re-ranking principal based on link evidence.

IV.2. Approaches Studying of Structure and Nature ofLinks in Wikipedia

The research concerning analysis and study of the nature of links in Wikipedia corpus aims mainly atidentifying the relationship between links and relevance. These works have analysed the structure of XML documents extracted from the INEX 2007 and INEX 2009 collections and their differences with Web link structure.

Junte Zhang and Jaap Kamps (2008) investigate and analyse some link related parameters, i.e. link density and repetition, document's length, the distance between anchor text occurrences, and the frequency of the anchor text within an article. Authors suppose that link repetition issue is directly related to link density and consider that this issue is still a challenge in the automatic link detection task. They use the INEX 2007 LTW (Link-The-Wiki) data to investigate and analyse these link related parameters. Figure IV.1, represents a distribution plot for all link occurrences of the un-orphaned sub-set of XML documents demonstrates that most of the links occurs once.

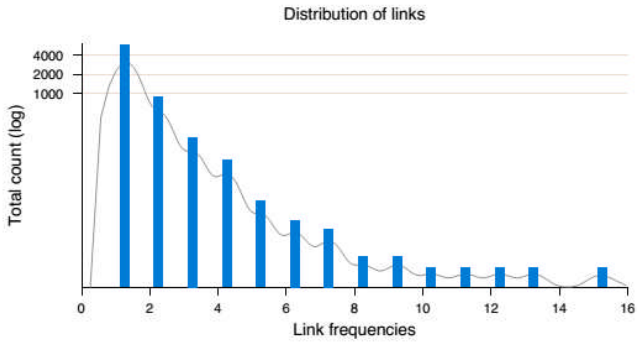


Figure IV.1: Distribution of all link frequencies (Zhang & Kamps, 2008)

Authors(Zhang & Kamps, 2008)noticed also that link density (number of appearing links) in the set of the un-orphaned documents is mostly consistent and dependent on the length of these documents (see Figure IV.2).

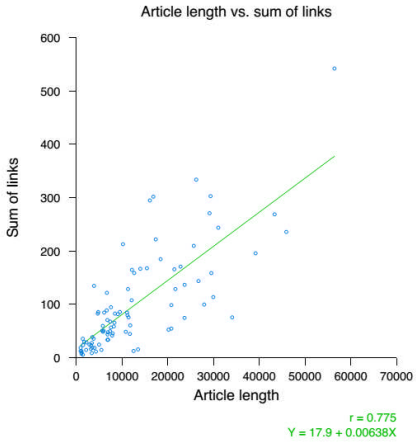


Figure IV.2: Strong positive correlation between article length and number of links (Zhang & Kamps, 2008)

Theyconducted experiments to consider the way the repeated linkscan be exploited. They found that even if the impact of link repetition is modest, performance can increase by using an informed approach to link repetition.

Figure IV.3 demonstrates that the majority of the INEX 2007 LTW topics have a similar link density.

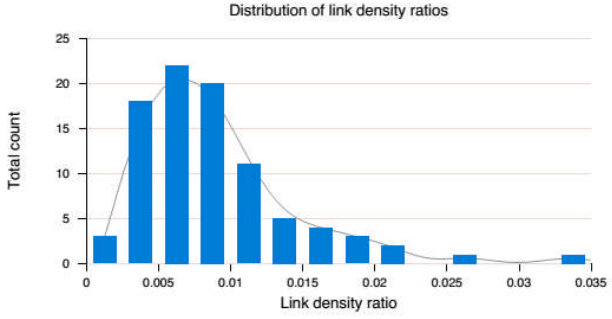


Figure IV.3: Distribution of link density ratios of 90 topics (Zhang & Kamps, 2008)

Junte Zhang and Jaap Kamps observed that the probability that a link is repeated is higher when either the anchor distance is shorter, or the number of repeated candidate links is smaller (see Figure IV.4).

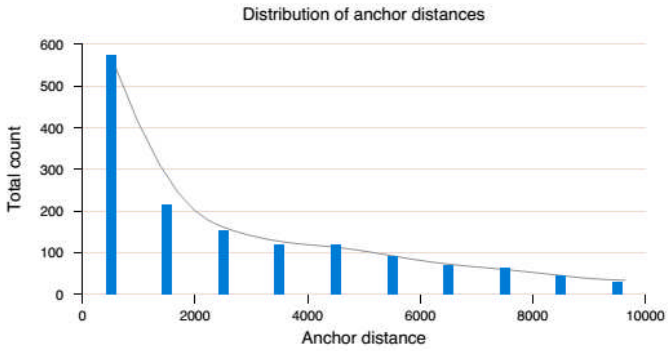


Figure IV.4: Distribution of anchor distances for an anchor a (Zhang & Kamps, 2008)

Authors use these statistics (link density, link repetition) to study their impact on the detection of links in LTW Track. They clearly show that the metrics used by INEX does not consider these structure properties.

By using repeated links authors (Zhang & Kamps, 2008)obtained higher precisions, which means that the detected links are well placed in terms of frequency and density. Also, they observe that taking into account the anchor distance improves the link detection performance (Run 2 ofTable IV.1).

Run	Precision	Recall	F-Score
Baseline 1	0.8459	0.8043	0.8206
Baseline 2	0.7526	0.9967	0.8053
Run 1 (AD)	0.7635	0.8750	0.7790
Run 2 (AD)	0.8517	0.8126	0.8279
Run 1 (RCL)	0.8445	0.8286	0.8295
Run 2 (RCL)	0.8517	0.8126	0.8279

TableIV.1 : Overall results for link detection(Zhang & Kamps, 2008)

Authors illustrate that there is a strong relation between link detection and the topicality of documents.

Marijn Koolen et al. (Koolen, Kazai, & Craswell, 2009) investigate the use of link structure of Wikipedia as an estimation of a book's relevance to the query in the Book Search track of INEX. Authors aim at incorporating additional sources of evidence extracted from Wikipedia to improve the retrieval effectiveness of a book search engine, for instance, exploiting the Wikipedia link structure to connect users' queries directly with relevant books.

The proposed method is motivated by the observation that Wikipedia collection often contain references to other information sources, such as: web pages, journal articles, books on the topic of the article. It consists at using the Wikipedia articles citations to retrieving books related to the user's query. Authors consider that if the topic of a Wikipedia article is relevant to the user's query then the cited books in that article can be considered as relevant.

They conclude that link distance between query and book pages in Wikipedia provides a good indicator of relevance that can boost the retrieval score of relevant books in the result ranking of a book search engine.

Jaap Kamps and Marijn Koolen (Kamps & Koolen, 2008) analyse the relation between link evidence and the relevance of pages in INEX 2007 Wikipedia collection. Theyshowed that the Wikipedia link structure can be considered as a "possibly weak" indicator of relevance and that significant improvement of retrieval effectiveness has been made with the local context. In the global context, they analyse incoming links to 5646 "relevant" XML documents and observe (Figure IV.5) that there is no absolute evidence in the link indegree (incoming links), i.e. both low indegree and high indegree pages (XML documents) can be relevant.Jaap Kamps and Marijn Koolen (Kamps & Koolen, 2008)zoom on the use of link evidence in local context by using the number of incoming links among a subset of local pages for a given topic. They observe also that local indegree is no absolute evidence of relevance.

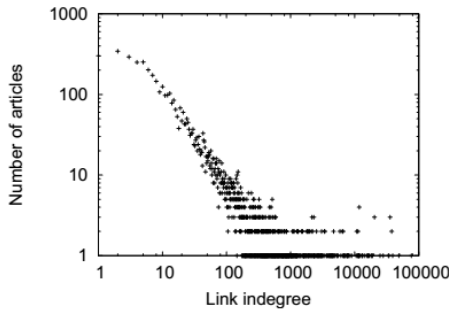


Figure IV.5: Wikipedia collection link indegree distribution of 5,646 "relevant" pages (Kamps & Koolen, 2008)

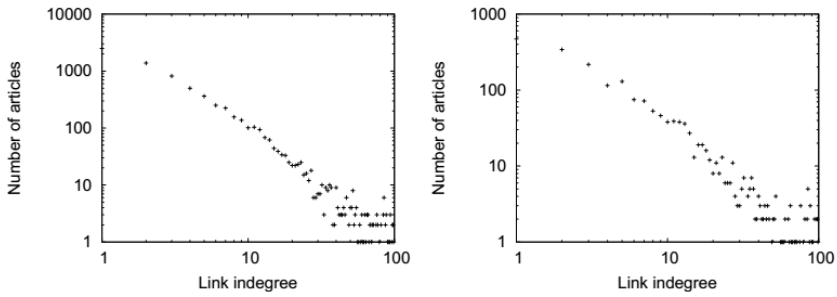


Figure IV.6: Wikipedia local link indegree distribution of 11339 local pages (left) and 2489 local relevant pages (right) (Kamps & Koolen, 2008)

Jaap Kamps and Marijn Koolen (Kamps & Koolen, 2009) investigate the difference of the link structure between Wikipedia and Web. For that they study structural aspects of two IR collections, i.e. the *.GOV* collection used at the TREC Web tracks (consisting of about 1.2 million documents) and the *Wikipedia* XML Corpus used at INEX 2007 (containing over 659 thousands XML articles). The main research question of their paper is to find out if, and how, the link structure of Wikipedia differs from the Web. They considered the three following issues:

• What is the distribution degree of links in Wikipedia and.GOV collections?

• Are there differences between distributions of incoming (indegree) and outgoing (outdegree) links?

• And, in particular, how does the link topology relate to the relevance of retrieval results?

Theyfirst performed a comparative analysis of Wikipedia and .GOV link structure and then consider the value of link evidence for improving retrieval accuracy. Table IV.2 provides statistics on the incoming and outgoing links and document lengths of these collections.

		min	max	mean	median	stdev
GOV	Indegree	0	44,228	8.90	1	126.00
	Outdegree	0	653	8.90	4	16.61
	Length	2	102,069	6,345	1,892	13,377
Wiki	Indegree	0	74,937	20.63	4	282.94
	Outdegree	0	5,098	20.63	12	36.70
	Length	16	281,150	2,473	1,309	4,238

TableIV.2 : Statistics of the .GOV and Wikipedia collections(Kamps & Koolen, 2009)

An author of a web page can link his page to any other page in the Web context, whether a topical relation exists or not between the two Web-pages. Whereas, in Wikipedia, links are based on words naturally occurring in a page and link to another related Wikipedia page. These links tend to be relevant to the local context. Authors noticed the existing of 1,269,988 (11.4%) reciprocal links in the .GOV collection, and 1,182,558 (8,7%) reciprocal links in the Wikipedia XML collection. They found that the average number of incoming and outgoing per document is 8.90 in the .GOV collection and 20.63 in Wikipedia (the maximal of outgoing links in Wikipedia was 5,098 which is much higher than 653 outgoing links of the .GOV collection).

The Authors observe smoother distributions and little difference between the incoming and outgoing links in Wikipedia compared to .GOV. Which means that outgoing links in Wikipedia behave very much like incoming links. They noticed that both indegree and outdegree seem to be good indicators of relevance. Another type of evidence that have been studied in (Kamps & Koolen, 2009)is the document length property. Authors observed that there is no evidence for the value of document length as indicator of relevance for the .GOV. Otherwise, study suggests that document length property of Wikipedia collection can be used as indicator of relevance.

The main findings of their study are, first, link structure of Wikipedia collection is similar to that of the Web, but more densely linked. Second, outgoing links of Wikipedia collection behave similar to incoming links, and both can be considered as good indicators of relevance, whereas in the Web context the incoming links are more important compared to outgoing links.

IV.3. Link Detection Approaches

Many IR methods have been proposed to construct hypertext on the Web (Agosti, Crestani, & Melucci, 1997; Allan, 1997; M. Henzinger, 2005) automatically, as well for determining missing links in Wikipedia Collection (Adafre & de Rijke, 2005; Fachry, Kamps, Koolen, & Zhang, 2008). Link detection is a specific case of the focused retrieval in which likely links between documents have to be identified automatically.Since 2007 a specific link-detection task at INEX campaign called Link-the-Wiki (LTW) has been defined. As defined at INEX's Link-the-Wiki track (Huang, Xu, Trotman, & Geva, 2008), the track focuses on automatically

linking an orphan page (an XML document denoted as topic file) to already existing Wikipedia pages (outgoing links; out-links) and from already existing Wikipedia pages to the orphan document (incoming links; inlinks). The file-to-file link discovery task aims at to analysing the textual content of a given Wikipedia document and recommend a set of up to 250 incoming and 250 outgoing links from and to other documents in the collection. Automatically detection of links is an important research topic because links are a fundamental feature of hypertext to navigate document collections. The task of creating links is a tough task since it requires a huge effort to decide which text fragments are important enough for the reader to serve as link anchor, and which documents are good targets for that text fragments. Additionally, determining the appropriate link targets implies knowledge of the complete collection, something which is hard to reach.

XML documents of INEX Wikipedia collection contain several types of links. These links have been implemented using XLink notation. Occurrences of these types of links in the un-orphaned (original) XML documents used in INEX 2007 LTW topics are presented in the TableIV.3. INEX LTW Task focuses on structural links, which have an anchor and refers to the Best Entry Point of another page (element level). Locally links, mainly used to improve navigation inside XML documents, was outside the scope of the INEX LTW track.

			All		Link in Article	
Type	**Uniq**	**Total**	**1×**	**Max**	**1×**	**Max**
`<collectionlink>`	5,786	8,868	4,275	51	5,781	15
`<unknownlink>`	1,308	1,458	1,201	14	1,271	7
`<outsidelink>`	807	851	772	5	778	5
`<imagelink>`	197	212	195	15	197	15
`<languagelink>`	79	1,147	12	66	1,147	1
`<wikipedialink>`	59	60	58	2	58	1
`<weblink>`	27	28	26	2	26	2
Total	8,263	12,624	6,513	-	9,232	-

TableIV.3 : Statistics of types of links in INEX LTW un-orphaned articles (Fachry et al., 2008)

We can observe from this table that the <*languagelink*> link type has 79 different forms (appearing once) used 1147 times. In the INEX Wikipedia collection a language link appears only once in a file. In the INEX LTW task, only three types of links are considered for detection: <*collectionlink*>, <*wikipedialink*> and <*unknownlink*>. The <*collectionlink*> link type includes of the majority of the links in the orphaned XML documents (70.0%).

The two most well-known link detection algorithms are those proposed by Kelly Itakura (Itakura & Clarke, 2008) and Shlomo Geva (Geva, 2008). The first algorithm chooses anchors in a new document by calculating the probability (γ) that each text fragment (phrase), if found in the already-linked part of the corpus, would be an anchor (Itakura & Clarke, 2008). If γ probability exceeds a certain threshold (which may be based on the length of the document), the text fragment is used as an anchor. The link target for is chosen to be the most common target for that anchor among existing links. The γ probability for a given text fragment *P*is defined by the following formula:

$$\gamma = \frac{number\ of\ occurences\ of\ P\ in\ the\ corpus\ as\ link}{number\ of\ occurences\ of\ P\ in\ the\ corpus\ altogether} \qquad \text{(IV.1)}$$

The second algorithm (proposed by Geva) searches in the text of the orphaned documents for occurrences of the corpus document titles. If such an occurrence is found, it will be used as an anchor. The target of the link is the document whose title was found.

In INEX 2007, existing links in origin XML nodes were removed from the 90 topics, which makes these XML documents 'orphans'. The aim of the LTW task was to detect these links again and find the correct XML node target (denoted 'fosters'), thus detecting both levels of links: element and document. More details can be found in the Appendix.

IV.4. Re-Ranking with Link Evidence Approaches

Link-based ranking algorithms exploit the link structure to determine the importance of nodes (elements or documents) in a link graph. The best-known algorithms are: degree-based (indegree/outdegree) and propagational algorithms (Pagerank, HITS, etc.)(Koolen, 2011).

As aforementioned only few works, exploiting XML links, have been conducted in the context of XML information retrieval.

Guo et al. (Guo et al., 2003) proposed XRANK, one of the first works which uses the links as source of evidence to compute the scores of XML elements. The link score (ElemRank) is based on three types of links between XML nodes, namely, CE links, which represent the internal links, HE links, which represent the hyperlinks (external links) and CE^{-1}, which represents the reversed internal links. XRANK architecture is described by Figure IV.7.

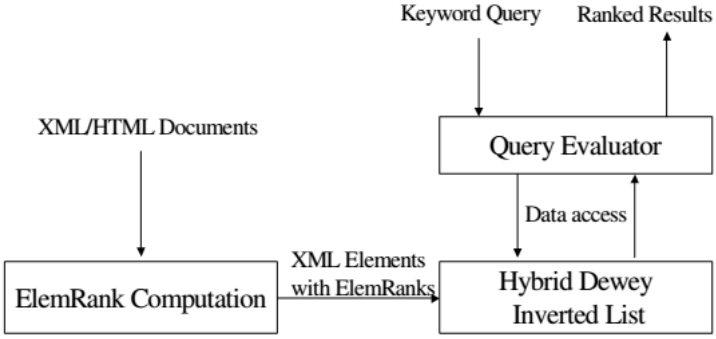


Figure IV.7: XRANK Architecture (Guo et al., 2003)

The link score computation module (ElemRank) is defined in as a measure of the objective importance of an XML element. This measure is computed based on the link graph structure of XML documents collection (global context). ElemRank algorithm is similar to PageRank algorithm, but is computed at the element granularity and by taking into account the nested structure of XML.They argue the use of bi-directional internal links (CE and CE^{-1} links) by the following example: "if a workshop contains many papers that have high ElemRanks, then the workshop should also have a high ElemRank", which corresponds to reverse ElemRank propagation.

They propose a formula based on three probabilities of navigating: d_1 (through hyperlinks), d_2 (forward containment edges), and d_3 (reverse containment edges).

$$e(v) = \frac{1-d_1-d_2-d_3}{N_d \times N_{de}(v)} + d_1 \sum_{(u,v)\in HE} \frac{e(u)}{N_h(u)} + d_2 \sum_{(u,v)\in CE} \frac{e(u)}{N_c(u)} + d_3 \sum_{(u,v)\in CE^{-1}} e(u \qquad \text{(IV.2)}$$

Where:

- e(v) denote the ElemRank of an element v.
- N_d is the total number of documents
- $N_h(v)$ is the number of outgoing hyperlinks from document v
- $N_{de}(v)$ is the number of elements in the XML documents containing the element v.
- $N_c(u)$ is the number of sub-elements of u

They performed their ElemRank computation algorithm on DBLP (143MB) and XMark (113MB) (Schmidt et al., 2002) by setting the d_1, d_2 and d_3 parameters to: 0.35, 0.25 and 0.25 respectively. The convergence threshold is set to 0.00002. The computation process for the entire dataset converged within 15 minutes (for both datasets) which suggests according to authors that computing ElemRanks at the elements granularity can be done offline in a reasonably time for large XML document datasets.

XRANK suffers from several negative points: first, the proposed link score computation formula is applicable exclusively in the global context (the entire collection context), whereas, all recent experiments that have been achieved on the INEX XML Wikipedia collection(Fachry et al., 2008; Kamps & Koolen, 2008; Kimelfeld et al., 2007; M'hamed Mataoui, Mezghiche, & Boughanem, 2010)showed that the use of XML links in the global context does not improve the retrieval accuracy; second, in the topical context we cannot ensure that all returned XML elements may have direct internal links with other retrieved XML elements; which means that XRANK cannot be adapted to topical (query-dependent) context because of the definition of the CE and CE^{-1} sets, i.e., these sets represent internal (hierarchical) links; third, many of XML IR tasks do not permit overlapping of returned XML elements, which make no sense to the proposed formula when applied to these IR tasks, i.e. an xml element and its parent cannot be returned in the same retrieval list; finally, XRANK method was experimented on two datasets: XMARK(Schmidt et al., 2002) and DBLP(DBLP). These two relatively small datasets are more suitable to the databases field than the information retrieval field because of their structure and content. The only experiment performed on the XRANK system is related to the performance, i.e., execution time, and not to the relevance of retrieval results.

Most of the works that come after XRANK appeared in the context of the INEX initiative.

In (Ramírez, Westerveld, & de Vries, 2005), authors presented an analysis of links between retrieved and relevant elements and exploited the findings to improve retrieval effectiveness in INEX 2005 Focused task. Experiments showed that the proposed approach outperforms the baselines presented in all settings. They indicated that the links discovered are very good pointers to highly relevant information. Authors showed also that using the maximum score of all linked elements gives better results than taking an average. This indicates that a single good linked element can already indicates the merit of this element.

Kimelfeld et al.(Kimelfeld et al., 2007) apply the HITS algorithm on the top-N returned documents to filter the results returned to the user. Authors study how to use the Wikipedia structure (XML documents with hyperlinks) by experimenting different combinations of language models and the HITS algorithm. They propose an approach to estimate the XML elements relevance consisting of two main steps:

- Identify a relevant subset of XML documents by using the F_{LM} filter based on statistical language modelling combined with smoothing techniques;
- Rank XML elements belonging to these XML documents (first step) by using the F_{HITS} filter based on an analysis of the links using the HITS algorithm.

To apply the HITS algorithm, authors define the link graph where the nodes represent the XML documents retrieved as response to a query q. This link graph is constructed by the following steps:

- Construct the set S_q of all documents D, such that a link to D contains one or more terms of q.
- Apply the filter F_{LM} to S_q and let S_{fq} be the set of the top-5 documents of S_q.
- S includes all the documents of S_{fq}, all documents that point to a document of S_{fq}, and all documents that are pointed to by a document of S_{fq}.

Authors apply element rankers (R_{LM} and R_{HITS}) to the XML elements in the filtered corpus to obtain the final rank. R_{LM} ranker is based on statistical language modelling and R_{HITS} combines R_{LM} ranker with the HITS rank of the XML document to which the element belongs as described in the following formula.

$$R_{HITS}(E) = R_{LM}(E) \cdot HITS(D), \tag{IV.3}$$

Where D is the document to which the XML element E belongs.

In their submissions to the INEX 2006, several combinations of filters, rankers and parameters were used.

			nxCG[5]		nxCG[10]		nxCG[25]		nxCG[50]	
runID	**method**	**co/s**	**Score**	**rnk**	**Score**	**rnk**	**Score**	**rnk**	**Score**	**rnk**
15_12_28	F_{HITS}/R_{LM}	co	0.4066	4	0.3827	2	0.3312	1	0.2770	2
16_08_52	F_{HITS}/R_{LM}	cos	0.3890	6	0.3697	4	0.3302	2	0.2816	1
16_12_44	F_{HITS}/R_{HITS}	cos	0.3999	5	0.3626	8	0.3152	5	0.2660	3
18_09_38	F_{LM}/R_{LM}	co	0.3878	7	0.3670	5	0.3163	4	0.2620	5
18_12_32	F_{LM}/R_{LM}	cos	0.3684	12	0.3506	9	0.3081	7	0.2639	4

TableIV.4 : Results in Focused task, with overlap off (Kimelfeld et al., 2007)

The results obtained using the HITS algorithm have not been convincing and the authors have proposed as prospects to use Pagerank instead of HITS. Mataoui et al. (M'hamed Mataoui et al., 2010) have also shown that using HITS algorithm on INEX 2007 dataset does not improve the retrieval performance at all.

Fachry et al.(Fachry et al., 2008), Kamps J. and Koolen M.(Kamps & Koolen, 2008) exploited the XML links to re-rank the retrieval results according to two levels: "global indegree" and "local indegree". The first level represents the number of incoming links to an XML document from the global collection and the second level represents the number of incoming links to an XML document from the documents returned as results for a given topic.

Authors investigated the effectiveness of incorporating link evidence as an indicator of relevance into their retrieval model in INEX campaign. They observe that link priors improve most of submitted runs for the Relevant in Context and Best in Context Tasks.

Instead of incorporating link evidence directly into the retrieval model, authors have chosen to use their priors to re-rank the returned elements.

They have only used the number of incoming links (indegree factor) per article. This factoris considered only at the article level. All the retrieved XML elements belonging to an XML document D will be multiplied with the same prior score.

$$Score = Score_{retrieved} \cdot Prior \tag{IV.4}$$

Two levels of incoming links are used: first, the global indegree (the number of incoming links from the entire collection), second, the local indegree (the number of incoming links from the documents returned as results for a given topic). The global indegree (local indegree) prior is proportional to the global degree (local degree) of an XML document respectively:

$$P_{\mathsf{Glob}} \propto 1 + global \tag{IV.5}$$

$$P_{\mathsf{Loc}} \propto 1 + local \tag{IV.6}$$

Otherwise, log based formulas have been used:

$$P_{\mathsf{LogGlob}} \propto 1 + \log(1 + global) \tag{IV.7}$$

$$P_{\mathsf{LogLoc}} \propto 1 + \log(1 + local) \tag{IV.8}$$

Also, authors used a third prior based on a weighted combination of the two first priors (noted $P_{\mathrm{LocGlob}}(d)$).

$$P_{\mathsf{LocGlob}} \propto 1 + \frac{local}{1 + global} \tag{IV.9}$$

The results obtained by using global indegree are not effective. Both the global indegree prior and the log global indegree prior lead to loss of performance for all three tasks.

By using the local link evidence, authors observed a loss of retrieval accuracy for Focused task (-4%) and a small improvement (about 3%) by using the (log) local indegree prior.

Experiment with the combined prior for the Focused Task obtained 8% of accuracy improvement. Authors observed that the combined local/global prior seems to be effective for improving ad hoc retrieval effectiveness.

Jaap Kamps and Marijn Koolen (Kamps & Koolen, 2008), used the same priors, with different combination (element, pool, contain). The combination "*element+link*" obtained the best interpolated precision at 0% recall. The proposed link prior shows a good improvement over the base run. They apply the link priors only to the first 100 retrieved XML documents per topic.

Run ID	Thorough MAep,off		Focused nxCG@10,off		Relevant in Context MAgP	
Baseline	0.0353		0.3364		0.1545	
Global Indegree	0.0267	-24.40***	0.1979	-41.16***	0.1073	-30.57***
Log Global Indegree	0.0335	-4.99	0.3066	-8.87**	0.1352	-12.50***
Local Indegree	0.0405	+14.75*	0.3218	-4.34	0.1467	-5.02*
Log Local Indegree	0.0418	+18.46***	0.3460	+2.85	0.1515	-1.96
Local/Global Indegree	**0.0463**	+31.08***	**0.3629**	+7.88**	**0.1576**	+1.99*

TableIV.5 : Results of using link evidence on three INEX 2006 ad hoc retrieval tasks. Best scores are in bold-face. Significance levels are 0.05(*), 0.01 (**), and 0.001 (***)(Fachry et al., 2008).

Authors observed that if the link evidence is made sensitive to the local context significant improvement of retrieval effectiveness is seen.

Run	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP
element	0.5672	0.4599	0.3137	0.2339	0.0707
element+link	**0.5999**	0.4745	0.3321	0.2753	0.0850
element+pool	0.5287	0.4705	0.3547	0.2729	0.0916
element+pool+link	0.5337	0.4779	0.3624	02938	0.1048
contain	0.5371	0.4728	0.3545	0.2952	0.0956
contain+link	0.5541	**0.4949**	0.3746	0.3156	0.1117
contain+pool	0.5289	0.4774	**0.3749**	0.2974	0.1011
contain+pool+link	0.5309	0.4821	0.3734	**0.3173**	**0.1157**

TableIV.6 : Results for the Ad Hoc Track Focused Task (Kamps & Koolen, 2008).

This approach is specific to document level granularity and can be misleading because, in general, the number of incoming links does not reflect the relevance of a document, but its link quality. This can be the case for a document pointed to by many links from irrelevant documents, instead of a document that can be pointed to by a single link from a highly relevant document.

Jovan Pehcevski et al. (Pehcevski, Vercoustre, & Thom, 2008) describe their approach of ranking entities extracted from the Wikipedia XML document collection. The proposed approach exploits the identified categories and the link structure of Wikipedia XML collection to improve the entity ranking effectiveness. Four principles constitute the proposed approach. Two of these principles relate to information extracted from links. First, as adaptation of the HITS, authors suppose that a page pointed to by a relevant page can be considered as an entity page. Second, a good entity page is pointed to by contexts with many occurrences of the entity examples.

Their experiments demonstrate that the locality of Wikipedia links can be exploited to significantly improve the effectiveness of entity ranking.

The proposed approach for entities ranking is based on a combination of three evidences: full-text similarity, page's categories similarity and links (by computing *linkrank*).

The *linkrank* formula computes a score for each XML document of the collection. This formula uses the number of incoming links to this XML document from the top-N results returned for a given query. The parameter *N* has been defined by experiments, by varying it between 5 and 100, the best performance is obtained with *N=20*. The proposed *linkrank* formula is:

$$S_L(t) = \sum_{r=1}^{N} \left(z(p_r) \cdot g(\#ent(p_r)) \cdot \sum_{l_t \in L(p_r,t)} f(l_t, c_r | c_r \in C(p_r)) \right) \quad \text{(IV.10)}$$

Where $S_L(t)$ represents the link score of the target entity t; $z(p_r)$ is the Zettair textual score.

The proposed weighting function $f(l_r,c_r)$ associated to the link l_t that belongs to the context c_r is defined as follows:

$$f(l_r, c_r) = \begin{cases} 1 & \text{if } c_r = p_r \text{ (the context is the full page)} \\ 1 + \#ent(c_r) & \text{if } c_r = e_r \text{ (the context is an XML element)} \end{cases} \quad \text{(IV.11)}$$

Where c_r is a context around entity examples that belongs to a set of contexts $C(p_r)$ found for the page p_r.

The combined (global) score $S(t)$ for a target entity page is computed as a linear combination of three scores: $S_L(t)$: the linkrank score; $S_C(t)$: the category similarity score; and $S_Z(t)$: the Zettair textual score:

$$S(t) = \alpha S_L(t) + \beta S_C(t) + (1 - \alpha - \beta) S_Z(t) \quad \text{(IV.12)}$$

α and β represent two parameters used for tuning according to the retrieval task.

The assumption adopted by Jovan Pehcevski et al.(Pehcevski et al., 2008)for considering locality of links is that references to entities (links) nearly located to the entity examples can be assessed as relevant.

They also observed consistent improvement in performance when names of the entity examples are added to the query.

	P[r]			
Run	5	10	R-prec	MAP
FullPage	0.1571	0.1571	0.1385	0.1314
StatL	0.2143	**0.2250**	**0.2285**	**0.1902**
StatR	**0.2214**	0.2143	0.2191	0.1853
DynCRE	**0.2214**	0.2107	0.2152	0.1828

(Q) Topic title

	P[r]			
Run	5	10	R-prec	MAP
FullPage	0.1857	0.1750	0.1587	0.1520
StatL	0.2429	0.2179	**0.2256**	0.2033
StatR	0.2429	**0.2214**	0.2248	**0.2042**
DynCRE	**0.2571**	0.2107	0.2207	0.1938

(QE) Topic title and entity examples

TableIV.7 : Performance scores for runs using different types of contexts in the *linkrank* module, obtained by different evaluation measures (Pehcevski et al., 2008).

All these link based approaches previously cited, except XRANK, do not propose solutions adapted to the "element-element" link type.

Philippe Mulhem and Delphine Verbyst (Mulhem & Verbyst, 2009; Verbyst & Mulhem, 2009) describe a method to incorporate link score in the computation of the final score of doxels (XML elements). Their approach is based on both exhaustivity and specificity scores between linked doxels. The proposed formula is applied in a global context. The adopted assumption of

their approach is that doxels are not only relevant because of their content but because they are linked to other relevant doxels.

They consider that the target doxel of a link is more exhaustive and/or more specific than the source doxel of this link.

They defined the properties of these measures (exhaustivity and specificity) based on formulas inspired by the overlap functions defined in(Salton & McGill, 1986).

$$Exh(d_1,d_2)=\frac{\sum_{i\in[1,n]|w_{2,i}\neq 0} w_{1,i}^{\ 2}}{\sum_{i\in[1,n]} w_{1,i}^{\ 2}} \quad \text{(IV.13)}$$

$$Spe(d_1,d_2)=\frac{\sum_{i\in[1,n]|w_{1,i}\neq 0} w_{2,i}^{\ 2}}{\sum_{i\in[1,n]} w_{2,i}^{\ 2}} \quad \text{(IV.14)}$$

The proposed RSV formula is a combination of two evidences (content and link evidences) by using the exhaustivity and specificity measures as a linear combination.

$$RSV_{env-link}(d,q)=\alpha.RSV_{\cos}(d,q) + (1-\alpha).\frac{1}{|env_{link}(d)|}.\sum_{d'\in env_{link}(d)}(\beta\, Exh(d,d')+(1-\beta)Spe(d,d'))RSV_{\cos}(d',q) \quad \text{(IV.15)}$$

Authors find that the approach with link environment obtained 2.2% of improvement for iP [0.00] and +17.0%, +21.5% and +21.7% for iP[0.01], iP[0.05] and iP[0.10] respectively. The obtained average accuracy is +10% by using link evidence. Authors showed, by experiments on the INEX XML collection, that "element-element" link type can improve retrieval accuracy.

IV.5. Conclusion

In this chapter, we have investigated the different proposed approaches for the use of link evidence in the XML information retrieval literature. We first classified these approaches into three categories: approaches thatanalyse the structure and nature of links in collections of XML documents; approaches proposing link detection strategies; approaches exploiting links to re-rank the list of elements initially returned by an information retrieval system. We focused in the context of our thesis on the last category of approaches.

The limitations of the proposed approaches are the following:

- The types of links used:Most of the proposed approaches only consider the navigational links; the hierarchical links are generally ignored (excepting for XRANK), while these links carries information that can be used effectively in the context of XML information retrieval.Moreover, in the literature there is almost no approach (except (Verbyst & Mulhem, 2009) which allows to take into account the case of "*element-element*" links.
- Known phenomena related to the algorithm used:Some work (Kimelfeld et al., 2007) proposed to adapt the HITS algorithm. This algorithm is well-known by the TKC phenomenon, which has led to lower results in the context of the structured IR.

- Context related issues: Several approaches propose to adapt the formulas in the global context (XRANK, global indegree, PageRank, etc.) (Guo et al., 2003; Kamps & Koolen, 2008), while it has been proved in several experiments that the *"topic-sensitive"* approaches are the most effective in the INEX framework of XML IR experiments.
- Other Properties ignored: The proposed approaches consider links as "tout ou rien", and do not consider some properties in the use of information relating, such as link distance, weight of link, link text, link type, etc. We believe that thisinformation can move forward reasoning on using links.

The next chapter consists in presenting all the proposed approaches to the exploitation of the links in the context of XML information retrieval.

Chapter V.

Harnessing Links in XML IR:the Proposed Approaches

V.1. Introduction

In this chapter, we will detail the approacheswe proposed to use link evidence in the XML information retrieval.We have divided the content of this chapter into 3 parts:

We start our statements by a statistical study that allowed us to observe that link evidence can play a role of relevance indicator.

In the second part of this chapter, we describe the way we calculate the link score of retrieved XML elements. Thus, we detail our three formulasof link Score computation: adapting Web link analysis algorithms to the XML context, distance based approach and weightedlinks based approach.

In the third part of this chapter, we describe the way we integrated the computed link score (or rank) in calculating the final score through three combination methods: Linear formula, Dempster-Shafer formula and Fuzzy formula.

Throughout this chapter, we will handle link evidence on two levels:

- *Global context.*At this level, we consider the entire link graph of the collection.
- *Local context.* At this level, we consider a subset of the collection as the link graph, by taking into account a subset containing the top ranked documents retrieved for a given query.

V.2. Some Statistics

To better understand the interest of using XML links as a source of evidence in the context of XML information retrieval, we conducted a statistical study aiming at findingthe relationship between the relevance of the returned XML elements and links (incoming or outgoing from other returned XML elements).

This study was conductedbased on the retrieval results returned by the DALIAN retrieval system for the 107 CO topics of INEX 2007. We consider for each topic a subset of returned results (20, 50 and 100 top XML documents). We measure the percentage of incoming links from relevant XML elements and percentage of outgoing links pointing to relevant elements. The following table shows some of the results obtained by taking the retrieved XML elements belonging to the top 20 XML documents. Results shown in Table V.1 can be read as follows: for instance,in topic 419 we have 40 incoming links to the relevant XML elements of top 20 retrieved elements, for which 33 links come from relevant elements. Also, for outgoing links, we have 35 outgoing links from elements of top 20 documents (to which retrieved relevant elements belong)that have 33 links pointing relevant elements.

For most queries, the probability that a relevant element is pointed-to by the relevant elements is greater than 50%. The same statement applies to the case of outgoing links. Which means that there is a close relationship between the relevance of the XML elements and links connecting these elements. In other words, a link from a relevant element (or pointing to a relevant element, respectively) can give an indication of the relevance of the target element (the source element, respectively).

	INLINKS			**OUTLINKS**		
TOPIC_ID	Relevant links	All links	% relevant	Relevant links	All links	% relevant
414	2	26	7,69	2	9	22,22
416	2	5	40,00	2	9	22,22
417	1	1	100,00	1	2	50,00
418	1	12	8,33	1	6	16,67
419	33	40	82,50	33	35	94,29
420	16	27	59,26	16	26	61,54
421	8	26	30,77	8	23	34,78
422	5	5	100,00	5	20	25,00
423	24	30	80,00	24	33	72,73
424	55	59	93,22	55	56	98,21
425	100	128	78,13	100	106	94,34
426	21	40	52,50	21	27	77,78
427	5	22	22,73	5	27	18,52
430	59	64	92,19	59	61	96,72
433	5	8	62,50	5	12	41,67
434	48	58	82,76	48	56	85,71
439	0	0	0,00	0	0	0,00
440	57	59	96,61	57	61	93,44
441	15	22	68,18	15	19	78,95
444	0	3	0,00	0	6	0,00
445	58	75	77,33	58	63	92,06
446	6	25	24,00	6	26	23,08
448	16	23	69,57	16	22	72,73
449	55	55	100,00	55	55	100,00
453	0	1	0,00	0	1	0,00
454	46	53	86,79	46	63	73,02
461	2	2	100,00	2	2	100,00
464	37	46	80,43	37	39	94,87
465	26	60	43,33	26	51	50,98
469	37	44	84,09	37	49	75,51
473	14	20	70,00	14	19	73,68
474	14	38	36,84	14	38	36,84
477	17	25	68,00	17	20	85,00
482	83	83	100,00	83	83	100,00
483	45	68	66,18	45	67	67,16
484	3	7	42,86	3	3	100,00
485	25	31	80,65	25	30	83,33

499	4	4	100,00	4	4	100,00
500	4	18	22,22	4	14	28,57
502	50	54	92,59	50	58	86,21
503	11	49	22,45	11	34	32,35
506	29	41	70,73	29	41	70,73
508	8	22	36,36	8	20	40,00
509	61	61	100,00	61	61	100,00
515	1	11	9,09	1	10	10,00
516	3	7	42,86	3	3	100,00
518	50	79	63,29	50	79	63,29
519	10	11	90,91	10	10	100,00
520	27	35	77,14	27	30	90,00
523	18	29	62,07	18	38	47,37
525	19	28	67,86	19	22	86,36
526	33	91	36,26	33	52	63,46
528	76	82	92,68	76	77	98,70
529	18	36	50,00	18	21	85,71
530	54	61	88,52	54	60	90,00
531	3	22	13,64	3	8	37,50
534	43	46	93,48	43	47	91,49
535	12	13	92,31	12	12	100,00
536	36	43	83,72	36	52	69,23
539	7	12	58,33	7	9	77,78
541	12	21	57,14	12	17	70,59
Total	1530	2167	70,60	1530	2004	76,35

Table V.1 : Percentage of link relevance (some of the 107 CO topics of INEX 2007, top 20 relevant documents)

All statistical tables containing the values for the 20, 50 and 100 top documents are listed in the Appendix I.

The following table (Table V.2) shows the results obtained by varying the number of relevant XML documents taken into account (top 20, 50 and 100) over the 107 CO topics of INEX 2007.We can note that the percentage of relevant links increases inversely with the number of top documents taken. This is also confirmed by the results obtained by application of our approach based on "*Topical_Pagerank*" (see section VI.4.1, tables: VI.1, VI.2 and VI.3)where the best results were obtained by using the top 20 XML documents.

	IN			OUT		
Top of Relevant documents	Relevant links	All links	% relevant	Relevant links	All links	% relevant
100	7165	14307	50,08	7165	11739	61,04
50	4529	8025	56,43	4529	6959	65,08
20	2130	3301	64,52	2130	2998	71,04

Table V.2 : Percentage of link relevance (some of the 107 CO topics of INEX 2007)

V.3. Proposition 1: Adapting Web Link Analysis Algorithms to the XML Context

The issueaddressed in this section can be defined by these two questions:

- Does the use of links as a source of evidence in the XML Information Retrieval context improves the quality of results, particularly in the case of the Wikipedia collection?
- Can the algorithms used by Web information retrieval be adapted to the XML IR context?

We examine how link evidence can be exploited in XML information retrieval (XML IR) field by experimenting some well-known link analysis algorithms, i.e., PageRank, HITS and SALSA.

XML IR systems incorporate both content and structure to compute similarity ($RSV(Q, E_i)$) between query Q and an XML element E_i according to retrieval models. We believe that this score can be combined with another score that represents link evidence computed according to one of the link analysis algorithms previously mentioned.The intuition behind our proposals is, if a document is referenced by several important documents in the XML collection then this may give an indication about its importance, the high importance of a document will therefore affect the scores of the returned elements in this document by an information retrieval system. The following figure shows a graph of links between a set of documents in the Wikipedia collection extracted as results to the topic 537 of INEX 2007 and a few links to documents that are not returned as results to that topic.

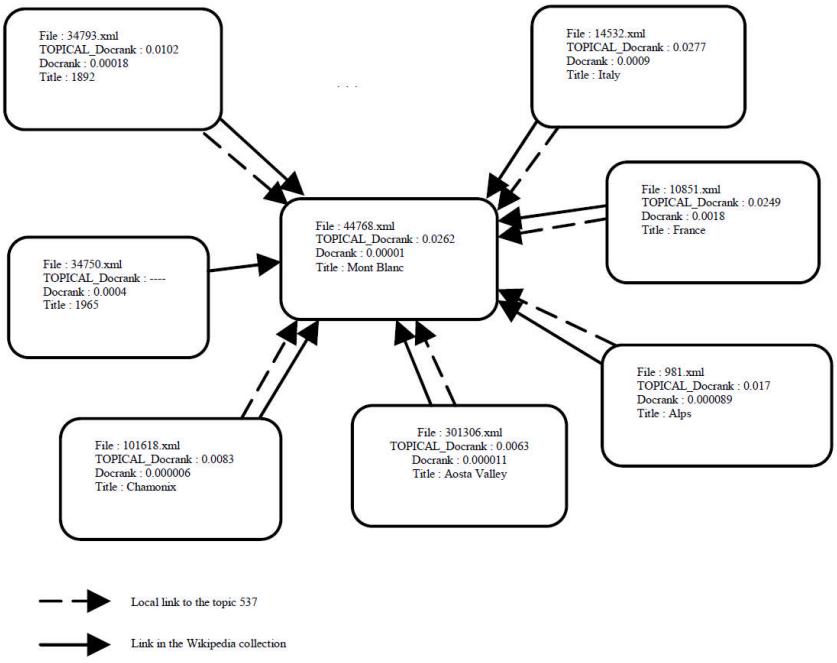


Figure V.1: Part of the links graph extracted from the INEX 2007Wikipedia collection with the "DOCRANK and TOPICAL_docrank" values computed for the "Topic 537".

This graph gives an idea about the structure of links between the documents in the Wikipedia collection which are of a semantic nature. In the topic 537 which has the title: "*pictures of Mont Blanc*" we note that a number of documents returned by the information retrieval system DALIAN point to the XML document "44768.xml" which has the title "*Mont Blanc*", which

reflects the high score affected by the application "*DOCRANK and TOPICAL_docrank*" formulas. If we introduce the score assigned to the XML document "44768.xml" it will increase its final score, and consequently the scores of retrieved XML elements that belong to it, which will improve the quality of results. In what follows we use the *"TOPICAL_docrank" and "TOPICAL_Pagerank"* designations to represent the same concept.

The first approach we proposed is based on an adaptation of PageRank to XML document collections. The "*DOCRANK*" of an XML document *D* in a collection of XML documents is computed according to the following formula:

$$DOCRANK(D) = \frac{(1-d)}{Nbr\_docs\_coll} + d * \sum_{(i,D)\in links} \frac{DOCRANK(i)}{|outlinks(i)|} \quad (V.1)$$

Where *links* represent all pairs *(i, D)* of links within the collection such that the document *i* contains a link to the document *D*. *Nbr_docs_coll* represent the number of documents in the collection and *d* represent the damping factor. |*oulinks(i)*| represents the number of outgoing links from the XML document *i*. The computation of "*DOCRANKs*" is done offline.

As aforementioned, the most important difference between global and local link evidence is that global evidence is query-independent while local evidence is query-dependent. For a document retrieved for two queries, the global link evidence will be the same, but the local link evidence is (probably) different for the two queries (Koolen, 2011).

Koolen (Koolen, 2011) observed that global link evidence is by nature query-independent, and is therefore not a direct indicator of the topical relevance of a document for a given query. As a result, he consider that link information is usually considered to be useful to identify the query-independent aspects of relevance referred as aspects of the importance of documents. The first assumption of Koolen is that global link evidence is not related to topical relevance but to document importance. He assumed that links represent a signal that linked documents which are topically related. Its second assumption is that local link evidence is related to topical relevance (in other words, local degree partly depends on the global degree and might also reflect the importance of documents).

In their analysis of the link graph of Wikipedia XML corpus Kamps & Koolen (Kamps & Koolen, 2008) showed that by zooming in on the local context of retrieved XML elements, i.e. the links between the top retrieved results, the number of incoming links can be used as an indicator of relevance to re-rank the result list.

They show also that incorporating link evidence in the retrieval model, for Wikipedia the global link evidence fails, so taking the local context into account is more effective (Kamps & Koolen, 2009). Furthermore, they conclude that the local degrees are still very effective for improving early precision, but are even more effective for general precision. Finally, they showed that the combined local/global evidence is less effective.

From All these observations, we have adapted the formula V.1 to the topical (local) context "*Query-dependent*", in which the computation is made with the same formula at the retrieval process time with a subset of returned results (*Nbr_docs_coll* parameter will represent the subset of returned documents used at computation). We refer to this new formula by : "*TOPICAL_docrank*". The computation of link scores is done in an iterative way following the same principle of PageRank to the convergence of "*DOCRANK*" or "*TOPICAL_docrank*" values.

V.4. Proposition 2: Distance Based Approach

In this second approach(M'hamed Mataoui & Mezghiche, 2015), we consider the problematic of efficiently generating ranked results in the XML IR context, by incorporating the link source of evidence. Despite of their popularity in the Web, only few research have exploited links to handle XML IR tasks. In contrast, we propose a new query-dependent link analysis approach based on a spreading-activation process that propagates relevance score through the two types of XML links, hierarchical and navigational, to compute a link score for each retrieved XML element. This propagation process depends on two features: the distance between elements and the type of the links separating these elements. The assigned link score will be combined with the content-based score to compute a new score used to re-rank the initial returned list of XML elements. Many features characterize our approach:

- First, it exploits element-to-document links to build element-to-element links for a given topic;
- Second, it exploits the two types of links: hierarchical and navigational links;
- Finally, it introduces the notion of "link distance" in the link score computation.

Contrary to the previous works(Fachry et al., 2008; Kamps & Koolen, 2008; Kimelfeld et al., 2007; Pehcevski et al., 2008) the approach we present attempts to exploit "element-element" links (path), composed either by internal (hierarchical) and/or external (navigational) links. Since most of XML collections contains "element-document" link type, we propose a solution that allows propagating "element-document" link to the elements of the target document.

V.4.1. Distance-Based Approach to Link Score Evaluation

We propose a query-dependent approach that combines content relevance score and link score to assign new relevance score to XML elements with respect to a query. This approach exploits these scores to re-rank elements initially retrieved by an XML retrieval system. Our focus concerns the way the links, both hierarchical (representing the internal structure of XML document) and navigational (representing external, element-document, links), are considered to assign link score to retrieved XML elements.

To compute this link score we define a collection of hyperlinked XML elements as a directed graph *G=(E, ELG, ILG)*; where *E*, the nodes of the graph, represents the set of retrieved XML elements; *ELG* represents the set of external links and *ILG* represents the set of internal links graph. We assume that internal links are bidirectional and external links are unidirectional. Therefore nodes are not all reachable from any node of *E*. We explore the idea, mainly exploited in web link-based algorithms, assuming that each retrieved element is assigned a given relevance score (which might be for instance its initial content score or any constant value). This relevance score is propagated between nodes through their links. It can be interpreted as the effort required to a user to navigate between the nodes of the link graph. So, the more the distance between two nodes is great, the more the effort to reach the target node is high. The remaining amount of relevance score actually received by the target is small. Therefore, the amount of relevance score that will be received by a given node is inversely proportional to its distance to the source node. Furthermore, we consider that the type of links play different roles when propagating this relevance score. We assume that for a user it is more profitable for him to navigate through external than internal links. Consequently, relevance score received by nodes from their internal links is of less interest than that received through external links.

Figure V.2, illustrates an example of link graph, which contains 3 XML documents: "*doc1.xml*", "*doc2.xml*" and "*doc3.xml*", containing the four XML elements, *N1*, *N2*, *N3* and *N4*, retrieved for a given query. These 3 documents are connected by 2 external (navigational) links. We notice that *N1* and *N2*can be reached from *N3* by navigating through external link *EL1*. *N4*can be reached from *N2* by navigating through external link *EL2*. *N1*can be reached from *N2* and *N2* from *N1* by navigating through hierarchical structure of "*doc1.xml*". Figure V.3, provides a simple graph representation of the example, where only retrieved elements are listed and their links weighted according to the distance (noted *d*) separating the corresponding nodes. The distance is seen as the number of edgesof the shortest path separating source and target nodes.

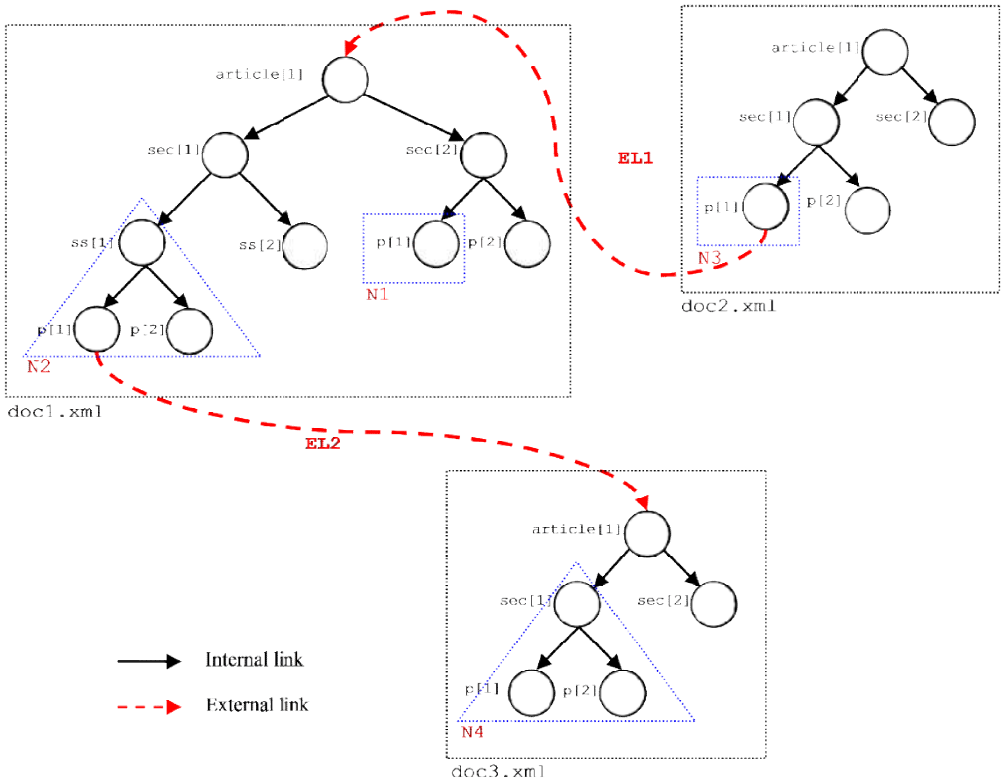


Figure V.2: Example of link structure graph (with internal and external links).

As aforementioned, we propose to compute a link score only for retrieved elements. Therefore, each retrieved element will propagate its relevance score toward its links to the other elements. The amount of propagated relevance score depends on the type of links (internal or external) and the distance separating the source node and its target nodes. The only constraint is that the total relevance score propagated by a given node cannot exceed (in our case we consider that it is equal) to its own relevance score. To formalize this process, let us consider *E(N)* the current relevance score of node *N*, and X_N as the unit of relevance propagated by node *N*. $E_{in}(N \rightarrow M)$ represents the relevance score received by node M from N. *dist(N,M)* represents the distance between *N* and *M*.

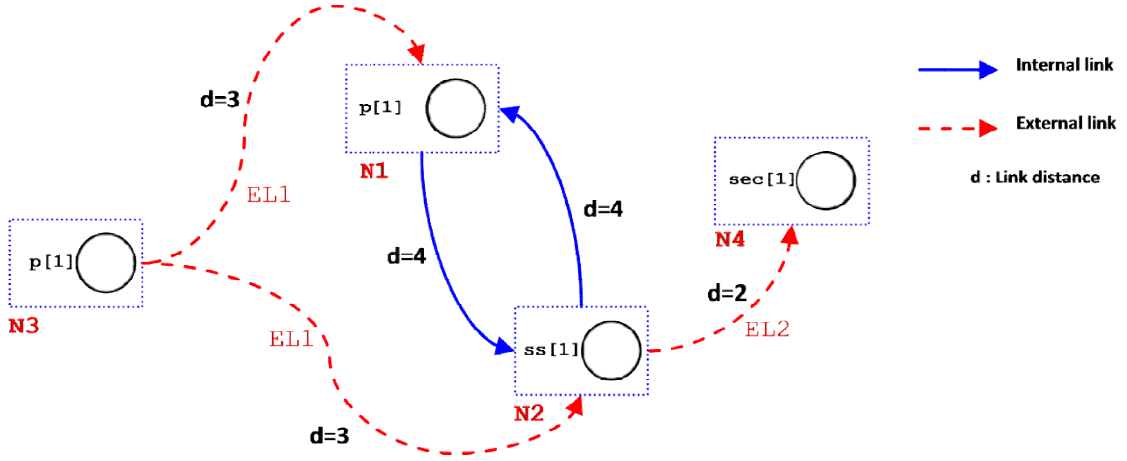


Figure V.3: Topical link graph with link distances.

To differentiate between relevance score transmitted through internal and external links, we propose to weight internal links relatively to external links. To formulate this idea we define a parameter β, and we replace X_N by $\beta * X_N$ in the case of internal links. The value of β parameter can be set to a value in the range from 0 to 1. The increase of β parameter value implies increasing of internal links importance.

The amount of relevance transmitted by a given node N to a target node M and the constraint associated with the total relevance propagated by a given node are as follows:

$$\begin{cases} E_{in}(N \to M) = \begin{cases} \dfrac{X_N}{dist(N \to M)} & if : (N \to M) \in ELG \\ \beta * \dfrac{X_N}{dist(N \to M)} & if : (N \to M) \in ILG \end{cases} & (V.2) \\ \sum_{M_i \in outlinks(N)} E_{in}(N \to M_i) = E(N) & (a) \end{cases}$$

Where:

- $E_{in}(N \to M)$ represents the relevance score propagated from N to M
- $E(N)$ represents the current link score of the XML element N
- X_N represents the spreaded relevance by N to an XML node M such as $dist(N \to M)=1$ (the value of X_N is computed for each XML node N, see equation V.4);
- $dist(N \to M)$ represents the distance that separate two XML elements: N and M;
- βis the parameter used to indicate the ratio between an internal link and an external link (the value of β is included in [0,1]).

Equation V.2.(a) represents the constraint about the total amount of relevance propagated by a given node which is equal to its own relevance score. Where M_i represents a node reached from outlinks of node N. We only consider active outlinks, i.e., those pointing retrieved elements.

We define two sets of XML nodes R_j and S_k such as: $(N \rightarrow R_j) \in ELG$ (External links graph); $(N \rightarrow S_k) \in ILG$ (Internal links graph) and $\{R_j \cup S_k\} = \{M_i\}$.

$$\sum_{M_i \in outlinks(N)} E_{in}(N \rightarrow M_i) = E(N) \tag{V.3}$$

$$\begin{aligned}
&\Rightarrow \left[\sum_{(N \rightarrow R_j) \in ELG} \frac{X_N}{dist(N \rightarrow R_j)} + \beta * \sum_{(N \rightarrow S_k) \in ILG} \frac{X_N}{dist(N \rightarrow S_k)} \right] = E(N) \\
&\Rightarrow X_N * \left[\sum_{(N \rightarrow R_j) \in ELG} \frac{1}{dist(N \rightarrow R_j)} + \beta * \sum_{(N \rightarrow S_k) \in ILG} \frac{1}{dist(N \rightarrow S_k)} \right] = E(N) \\
&\Rightarrow X_N = \frac{E(N)}{\left[\sum_{(N \rightarrow R_j) \in ELG} \frac{1}{dist(N \rightarrow R_j)} + \beta * \sum_{(N \rightarrow S_k) \in ILG} \frac{1}{dist(N \rightarrow S_k)} \right]}
\end{aligned} \tag{V.4}$$

By using equations V.2, V.3 and V.4, the link score of an XML element *P (LS(P))* is computed by summing relevance scores received from the different nodes, as follows:

$$LS(P) = \frac{(1-DF)}{NBR\_elements} + DF * \left[\sum_{N_r \in EL(P)} \frac{X_{Nr}}{dist(N_r \rightarrow P)} + \beta * \sum_{N_r \in IL(P)} \frac{X_{Nr}}{dist(N_r \rightarrow P)} \right] \tag{V.5}$$

With

$$X_{Nr} = \frac{E(N_r)}{\left[\sum_{(N_r \rightarrow R_j) \in ELG} \frac{1}{dist(N_r \rightarrow R_j)} + \beta * \sum_{(N_r \rightarrow S_k) \in ILG} \frac{1}{dist(N_r \rightarrow S_k)} \right]} \tag{V.6}$$

Where:

- *DF* represents the damping factor (usually set to 0.85)
- *NBR_elements* represents the number of XML nodes in the topical link graph
- $dist(N_r \rightarrow P)$ represents the distance between XML nodes N_r and *P*
- *EL(P)* represents the set of external links that points to *P* (external inlinks of *P*)
- *IL(P)* represents the set of internal links that points to *P* (internal inlinks of *P*)

"(1-DF)/NBR_elements" represents the probability of visiting an XML element *P* at random. The second part of the formula is the probability of visiting *P* by navigating through other XML elements by using both internal and external links. α and β can be tuned differently depending on the retrieval task.

Our link score formula is conceptually similar to Pagerank (used in the Web context), except that: **(a)** first, it is defined at the element granularity; **(b)** second, it takes into account the two types of links, i.e., hierarchical and navigational; **(c)** third, it is query dependent, i.e., applied in the topical context; **(d)** finally, it introduces a new feature, called link distance, to weight the amount relevance score propagated through XML links.

The computation of *LS(P)* is done in an iterative way until the convergence of link scores (Haveliwala, 1999). The convergence proof of our formula is similar to that described in(Farahat, LoFaro, Miller, Rae, & Ward, 2006).

V.4.2. Illustration

As example from Figure V.2, the propagated relevance from *N3* to *N1* and *N2* is formulated by: $[E_{in}(N3 \rightarrow N1) + E_{in}(N3 \rightarrow N2)] \leftarrow E(N3)$

$$\Rightarrow \frac{X_{N3}}{dist(N3, N1)} + \frac{X_{N3}}{dist(N3, N2)} \leftarrow E(N3)$$

$$\Rightarrow \frac{X_{N3}}{3} + \frac{X_{N3}}{3} \leftarrow E(N3) \Rightarrow X_{N3} \leftarrow \frac{3}{2} E(N3)$$

$$\Rightarrow E_{in}(N3 \rightarrow N1) \leftarrow \frac{E(N3)}{2} \quad \text{and} \quad E_{in}(N3 \rightarrow N2) \leftarrow \frac{E(N3)}{2}$$

This means that *N1* and *N2* will receive each one the half of relevance score propagated by *N3*, because the effort made by a user to reach *N1* or *N2* from *N3* is the same.

In the case of *N1*, two XML nodes can be reached: *N2* and *N4*. *N2*is reached via hierarchical structure and *N4* is reached via external link. The propagated relevance from *N1* to *N2* and *N4*is formulated by:

$[E_{in}(N1 \rightarrow N2) + E_{in}(N1 \rightarrow N4)] \leftarrow E(N1)$

$$\Rightarrow \frac{\beta * X_{N1}}{dist(N1, N2)} + \frac{X_{N1}}{dist(N1, N4)} \leftarrow E(N1)$$

$$\Rightarrow \frac{\beta * X_{N1}}{4} + \frac{X_{N1}}{2} \leftarrow E(N1) \Rightarrow X_{N1} \leftarrow \frac{4}{\beta + 2} E(N1)$$

$$\Rightarrow E_{in}(N1 \rightarrow N2) \leftarrow \frac{\beta}{\beta + 2} * E(N1) \quad \text{and} \quad E_{in}(N1 \rightarrow N4) \leftarrow \frac{2}{\beta + 2} * E(N1)$$

V.5. Proposition 3: Weighted Links Based Approach

We propose in this section another alternative (or another formulation) of our distance based approach(M'hamed Mataoui & Mezghiche, 2015). This Alternative approach(M'hamed Mataoui, 2014; M'hamed Mataoui, 2015), that we call "*Weighted Links Based Approach*", is a "topic-sensitive" approach that combines both initial content relevance score and link evidence score to compute a new relevance score for each retrieved XML element. We focused on the manner XML links, both navigational and hierarchical links could be used to compute link evidence score of retrieved XML elements.

To introduce the way the link score is computed we define a hyperlinked collection of XML elements returned as retrieval results for a given topic *Q* as a directed graph *Ω=(Q, E, NLTG, HLTG)*; where *Q* represents the topic (query) for which retrieved XML elements are returned as response; *E* represents the nodes of the graph, i.e., the set of retrieved XML elements in response to *Q*; *NLTG* represents the navigational (external) links and *HLTG* the hierarchical (internal) links between XML elements belonging to *E*. Navigational links are supposed as unidirectional links and hierarchical as bidirectional links. We explore principally the popularity propagation model exploited in web link analysis algorithms.

We assume that each retrieved XML element has a given relevance score that can be propagated through links. In our approach we interpret the amount of relevance score propagated between two XML elements, E_1 and E_2, as the probability to explore this path by a user. The propagated amount of relevance is inversely proportional to the *"path weight"*. Therefore, the more the path weight between two XML nodes is great, the more the probability to explore this path by a user is less. In our context a path consist of 0 or 1 navigational link and a set of hierarchical links. By considering that it is easier for a user to navigate through navigational (click on the link) than hierarchical links, we assume that the probability that a user traverses a path containing an navigational link is higher than that of a user traverses a path which contains only hierarchical links. Consequently, the propagated relevance depends on the existence of navigational link and the number of links. We call this concept: "*weighting of the links*", where we define a parameter λ that reflects the weight of navigational links (*NLW*) compared to hierarchical links (*HLW*). We propose the following formula:

$$NLW = \lambda * HLW \quad / \quad \lambda \in]0,1] \tag{V.7}$$

Increasing of λ value implies increasing of hierarchical links weight. The algorithm of computation of the path weight is shown in the algorithm of Table V.3. As aforementioned, we consider in our approach the two types of links: navigational links (*NL*) and hierarchical (*HL*). Navigational links connect generally between XML nodes belonging to different XML documents and hierarchical links represent the structure of these documents. As we have mentioned, our approach is applied in "topic-sensitive" context, which means that we exploit a sub-graph of the global link graph. This sub-graph can be obtained by incorporating two entities, which are: retrieval results and global link graph. To obtain the "topic-sensitive" link graph we extract the two link-type graphs as shown in Figure V.4 and Figure V.5.

if $\exists EP/(N_i \rightarrow EP)$ *is a navigational link* **and** $(N_i \rightarrow N_j) \equiv (N_i \rightarrow EP) \cup (EP \rightarrow N_j)$ **then**
  $PW(N_i,N_j) \leftarrow [NLW + [dist(EP,N_j) * HLW]]$
else
  $PW(N_i,N_j) \leftarrow [dist(EP,N_j) * HLW]$
end if

Table V.3 : Path weight "$PW(N_i,N_j)$" Computation Algorithm

To illustrate how "Path Weight" information is used to compute link scores of the retrieved XML elements we consider the example ofFigure V.4. Let a link graph containing four XML documents: "*document1.xml*", "*document2.xml*", "*document3.xml*" and "*document4.xml*". These documents contain five retrieved elements for a given query *Q*: *Node1*, *Node2*, *Node3*, *Node4* and *Node5*. These XML elements are connected by 3 navigational links *NL1*, *NL2* and *NL3*. We notice that *Node3* and *Node4*can be reached from *Node1* by traversing *NL1*. *Node3* and *Node4*can also be reached from *Node2* by traversing *NL2*. *Node5*can be reached from *Node3* by navigating through *NL3*. *Node3*can be reached from *Node4* and *Node4* from *Node3* by navigating through hierarchical structure of "*document3.xml*".

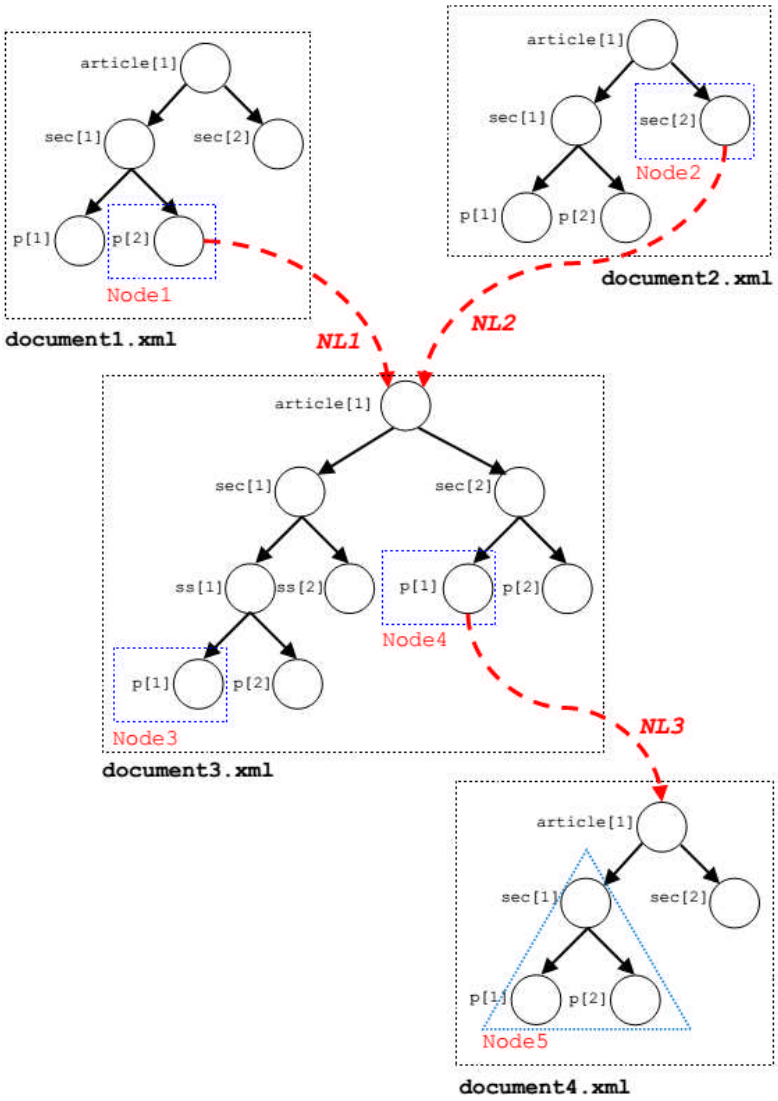


Figure V.4: Example of link structure graph (hierarchical and navigational links).

Figure V.5 represents a sub-graph of the link structure of Figure V.4 where only retrieved elements and their links weighted according to algorithm 1 are mentioned.

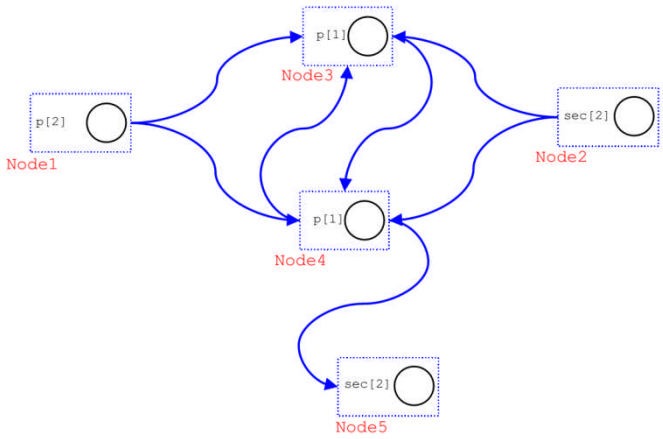


Figure V.5: "Topic-sensitive" link graph construction for the example of Figure V.4.

We now consider the problem of computing link scores of XML elements. As mentioned, the link score is a measure of the XML element importance, and it is computed based on the topic-sensitive link graph, i.e., retrieved XML elements. To compute the amount of propagated relevance score that passes through the link structure connecting two XML nodes N_i to N_j, we propose the formulas of equation V.8 taking into account the two types of links and the path weight between these XML elements.

To formalize this propagation process, we consider $RS(N_i)$ as the current relevance score of XML node N_i, and $URS(N_i)$ as the unit of propagated relevance score by N_i through a path with $PW=1$. $PRS(N_i \rightarrow N_j)$ represents the propagated relevance score by XML node N_i to N_j. $PW(N_i \rightarrow N_j)$ represents the weight of the path between N_i and N_j computed according to algorithm 1.

$$\begin{cases} PRS(N_i \rightarrow N_j) \leftarrow \dfrac{URS(N_i)}{PW(N_i \rightarrow N_j)} \\ \displaystyle\sum_{N_j \in outlinks(N_i)} PRS(N_i \rightarrow N_j) = RS(N_i) \end{cases} \tag{V.8}$$

Second part of equation V.8 represents the constraint related to the sum of the amount of relevance scores propagated by a given XML node which must not exceed (be equal) to its own relevance score. We define N_j as the set of XML nodes reached from outlinks of XML node N_i. Only active outlinks, i.e., those pointing to retrieved elements are considered. Equation V.10 represents the way the unit of propagated relevance score by N_i through a path (with $PW=1$) is computed.

$$\begin{aligned} &\sum_{N_j \in outlinks(N_i)} PRS(N_i \rightarrow N_j) = RS(N_i) \\ &\Rightarrow \sum_{N_j \in outlinks(N_i)} \frac{URS(N_i)}{PW(N_i \rightarrow N_j)} = RS(N_i) \end{aligned} \tag{V.9}$$

$$\Rightarrow URS(N_i) = \frac{RS(N_i)}{\displaystyle\sum_{N_j \in outlinks(N_i)} \frac{1}{PW(N_i \rightarrow N_j)}} \tag{V.10}$$

The final link score "*LS(XE)*" of an XML element *XE* is computed following equation V.11. "*LS(XE)*" is obtained by summing propagated relevance scores through different links, as follows:

$$LS(XE) = \frac{(1-\rho)}{|N|} + \left[\rho * \sum_{N_i \in inlinks(XE)} PRS(N_i \rightarrow XE) \right]$$

$$\Rightarrow LS(XE) = \frac{(1-\rho)}{|N|} + \left[\rho * \sum_{N_i \in inlinks(XE)} \frac{URS(N_i)}{PW(N_i \rightarrow XE)} \right]$$

$$\Rightarrow LS(XE) = \frac{(1-\rho)}{|N|} + \left[\rho * \sum_{N_i \in inlinks(XE)} \frac{URS(N_i)}{PW(N_i \rightarrow XE)} \right]$$

$$\Rightarrow LS(XE) = \frac{(1-\rho)}{|N|} + \left[\rho * \sum_{N_i \in inlinks(XE)} \frac{\frac{RS(N_i)}{\sum_{N_j \in oulinks(N_i)} \frac{1}{PW(N_i \rightarrow N_j)}}}{PW(N_i \rightarrow XE)} \right] \quad \text{(V.11)}$$

Where:

• $|N|$ represents the number of retrieved XML elements (nodes in the topic-sensitive link graph);

• ϱ parameter represents the damping factor (generally fixed at 0.85).

"$(1-\varrho)/|N|$" represents the probability of visiting randomly an XML element E in the graph of links. The second fragment of equation V.11 represents the probability of reaching E by navigating through both link types from other XML elements. Computation of *LS(XE)* is carried out according to an iterative process until the convergence of link scores.

V.6. Combination Formulas

In this section we detail the different combination formulas used to incorporate the link evidence score in the formula of the final score. We propose as part of our work to use three formulas: linear formula, Dempster-Shafer formula and Fuzzy-based formula. Both link score information and/or rank information can be used in the combination formula. We carried out experiments (see next chapter) about the impact of using combination formula based on "element rank" information instead of the scores. The obtained re-ranking results did not show any improvement of the retrieval accuracy but rather the contrary (diminution about -10%). We believe that part of the obtained improvement is due to the good choice of the combination formula as well as the used parameters.

To remove the effect of the large difference between "Link scores" and the "Initial scores", we require normalizing these scores before the computation of the "Final scores".

V.6.1. Linear Formula

The first approach we propose combines link score with initial score to obtain a final score of the returned XML elements(M'hamed Mataoui & Mezghiche, 2009, 2015; M'hamed Mataoui et al., 2010). Let the initial scores of the retrieved elements for a given query Q be: $IS(e_1)$, $IS(e_2)$, …, $IS(e_n)$ and the computed link scores be: $LS(e_1)$, $LS(e_2)$, …, $LS(e_n)$. The combined score is defined by:

$$FS(e_i) = f(IS(e_i), LS(e_i)) \quad \text{(V.12)}$$

Where:

- f is an aggregation function.

In most of our experiments we define f according to this linear combination formula (other formulas are supported, i.e., combined ranks):

$$FS(e_i) = \alpha * IS(e_i) + (1-\alpha) * LS(e_i) \quad \text{(V.13)}$$

Where:

- $FS(e_i)$ represents the final computed score for the XML element e_i;
- $IS(e_i)$ represents the initial score of e_i (obtained by application of XML retrieval model);
- $LS(e_i)$ represents the computed link score of e_i;
- α is a parameter that determines the degree of contribution of each score, it can be set to a value in the range [0,1].

This formula allow the re-rank of the initially returned list of XML elements according to their *IS* and *LS* scores.

V.6.2. Dempster-Shafer Formula

The second combination approach we propose is an evidential link-based approach for re-ranking XML retrieval results by using the Dempster-Shafer theory of evidence.It combines content relevance evidence for each retrieved XML element with its computed link evidence (score and rank). The use of the Dempster–Shafer theory is motivated by the need to improve retrieval accuracy by incorporating the uncertain nature of both bodies of evidence (content and link relevance). The link score iscomputed according to our new link analysis algorithm based on weighted links (see section V.5.), where relevance is propagated through the two types of links, i.e., hierarchical and navigational. The propagation, i.e. the amount of relevance score received by each retrieved XML element, depends on link weight defined according to two parameters: link type and link length (path weight).

Based on the well-known mathematical theory of Dempster–Shafer (also known as belief function theory),some approaches have been proposed in the literature. Lalmas and Ruthven (Lalmas & Ruthven, 1998) used the DS theory of evidence to combine aspects of information use. The proposed model combines evidence from user's relevance with algorithms describing how words are usedwithin documents. They also present some experimenting on this theory in information retrieval. Schocken and Hummel(Schocken & Hummel, 1993) used DS theory to combine taxonomies of keywords. In their approach different confidence levels are assigned for each defined keyword set. Then, using DS theory, they combine these assignments to find the new mass distribution over these sets. The use of this theory is mainly motivated by the incorporation of the uncertain nature of information retrieval.

In our context, we propose an evidential link-based approach for re-ranking XML retrieval results. The proposed approach is based on a combination of textual and structural information. Contrary to the previous works the approach we present attempts to exploit "element-element" links (path), composed either by internal (hierarchical) and/or external (navigational) links.Since most of XML collections contains "element-document" link type, we have proposed a solution that allows to propagate "element-document" link to the elements of the target document (see sectionV.5.).

Before describing the approach, we briefly define some notion of DS theory we exploited.

V.6.2.1. DS theory elements

The Dempster-Shafer (DS) theory (known as belief functions) theory is a theory of uncertainty that was developed by Dempster (Dempster, 1967)and further extended by Shafer(Shafer, 1976). This theory allows bettering quantifying uncertainty by allowing the explicit representation of ignorance, and it has attractive properties which provide significantly richer information in combining sources of evidence. This theory have been used to model various aspects of the information retrieval process(Lalmas & Ruthven, 1998; Schocken & Hummel, 1993).

The DS theory is based on the grounds of the following concepts and principles:

(a) The frame of discernment is a set of mutually exclusive and exhaustive hypotheses about the problem domains. From a frame of discernment (Θ) correspondingly 2^{Θ} is the power set of (Θ).

(b) A basic belief assignment (*bba*) or mass function represents the degree of belief and is defined as a mapping $m(\cdot)$ satisfying the following properties:

- $m(\emptyset) = 0$, $\emptyset$: the empty set

- $\sum_{H \in 2^{\Theta}} m(H) = 1$, H: a subset of Θ

The subsets H of the power set 2^{Θ} with a positive mass of belief is called focal set element of $m(\cdot)$.

(c) The Dempster's combination rule is the most important tool of the evidence theory. This rule aims to aggregate evidence from multiple independent sources defined within the same frame of discernment.

Let m_1 and m_2 be the mass functions associated with two independent bodies of evidence. H_1 and H_2 represent the focal elements of m_1 and m_2 respectively. The mass function m is formed by combining m_1 and m_2 as $m = m_1 \oplus m_2$. This rule with two sources, $m = m_1 \oplus m_2$ is defined by equation V.14.

$$m^{DS}(H) = \frac{m_{12}(H)}{1 - m_{12}(\phi)} \tag{V.14}$$

Where,

$$m_{12}(H) = \sum_{\substack{H_1, H_2 \in 2^{\Theta} \\ H_1 \cap H_2 = H}} m_1(H_1) m_2(H_2) \tag{V.15}$$

$m_{12}(H)$ and $m_{12}(\emptyset)$ represent the conventional conjunctive consensus operator and the conflict of the combination between the two sources respectively. Additionally, from a given *bbam*, the belief and the plausibility functions are used as decision criteria(Shafer, 1976).

V.6.2.2. The discounting of sources of evidence

It is possible to discount an unreliable source proportionally to its corresponding reliability factor according to the method proposed by Shafer(Shafer, 1976). Shafer assumes that if we know the reliability/confidence factor α that belong to the interval [0,1], then the discounting of the *bba m(·)* provided by the unreliable source denoted by *m′(·)* is defined as follows:

$$\begin{cases} m'(A) = \alpha.m(A), \quad \forall A \in 2^{\Theta}, A \neq \Theta \\ m'(\Theta) = (1-\alpha) + \alpha.m(\Theta) \end{cases} \tag{V.16}$$

V.6.2.3. Using the DS theory in XML Information Retrieval field

Within the context of information retrieval and according to the proposed new "topic-sensitive" approach, we define the frame of discernment by: $\Theta=\{e_i , \neg e_i\}$, where e_i is a retrieved element. Let S_1 and S_2 be initial (*Retrieval Model*) and link (*Our link score computation formula*) information retrieval sources respectively. Then, we define two basic belief assignments for initial and link scores obtained from S_1 and S_2 as follows:

- $m_{Si}(\emptyset)=0$, $\emptyset$: the empty set;
- $\sum_{H\in 2^{\Theta}} m_{S_i}(H)=1$, H : a subset of Θ and $S_i \in \{S_1 , S_2\}$initialand link scores (noted *IS* and *LS* respectively) can be scaled to fall between 0 and 1 in order to satisfy the mass properties as follows:

$$\begin{cases} m_{s_1}(e_i) = \dfrac{IS(e_i)}{\sum_{j=1...n}(IS(e_j))} \\ m_{s_2}(e_i) = \dfrac{LS(e_i)}{\sum_{j=1...n}(LS(e_j))} \end{cases} \tag{V.17}$$

Where *n* denotes the number of retrieved XML elements.

For XML elements classification decision making, we adopt the combination of initial and link information retrieval scores. This combination is based on Dempster's rule to obtain a final score mass of the returned XML elements.

Let the initial score masses of the retrieved elements for a given query Q be: $m_{S1}(e_1)$, $m_{S1}(e_2)$, $\cdots$, $m_{S1}(e_n)$ and the computed link score masses be: $m_{S2}(e_1)$, $m_{S2}(e_2)$, $\cdots$, $m_{S2}(e_n)$. Then, the combined score mass using Dempster's rule is definedby:

$$m^{DS}(FS) = m_{s_1}(IS) \oplus m_{s_2}(LS) \tag{V.18}$$

$$m^{DS}(FS(e_i)) = \frac{m_{12}(e_i)}{1-m_{12}(\phi)} \tag{V.19}$$

Where,

$$m_{12}(e_i) = \sum_{\substack{H_1,H_2 \in 2^{\Theta} \\ H_1 \cap H_2 = e_i}} m_{S_1}(H_1) m_{S_2}(H_2) \tag{V.20}$$

The preceding combination rule does not take into account the discounting factor of the two sources. To deal with the discount problem, we propose a novel discounting method, which can maximize for a given query, a scoring function that implicitly imposes an ordering on documents, directly defined on the rank performance measures. As a result, our discount approach uses a query-dependent ranking model to discount its score. According to each source, this method computes discounting factor of each element (e_i) on the basis of its rank because the ranking measure plays an important role in almost all activities related to information retrieval.

When a new query is consulted, the individual element rank in respect to the source S_j is obtained which is then used to compute the corresponding element discounting factor. This discounting factor is defined by the following formula.

$$\alpha_{S_j}(e_i) = \frac{1}{r_{S_j}(e_i)} \tag{V.21}$$

Where $r_{S_j}(e_i)$ denotes the rank of the element e_i according to their relevance to the query for the user in respect to the source S_j. Hence, using the Shafer's discounting of each source of evidence S_j and its corresponding factor $\alpha S_j(e_i)$, we proceed to calculate the reliability of each score mass of the element e_i which is defined as follows:

$$\begin{aligned} m'_{S_j}(e_i) &= \alpha_{S_j}(e_i) \cdot m_{S_j}(e_i) \\ m'_{S_j}(\neg e_i) &= \alpha_{S_j}(e_i) \cdot m_{S_j}(\neg e_i) \\ m'_{S_j}(\Theta) &= (1 - \alpha_{S_j}(e_i)) + \alpha_{S_j}(e_i).m_{S_j}(\Theta) \end{aligned} \tag{V.22}$$

Now for each element e_i, we apply Dempster's rule for combining their discounting initial and link scores. This is defined by the following equation:

$$m^{DS}_{(S_1,S_2)}(e_i) = m'_{S_1}(e_i) \oplus m'_{S_2}(e_i) \tag{V.23}$$

The final scores $m^{DS}_{(S_1,S_2)}(e_i)$ for $i = 1 \cdots n$ allow the re-rank of the initially returned list of XML elements based on DS theory that use the two "element-element" link types and fixed discounting rates according to the rank function of the elements.To show the utility and the effectiveness of these discounting rates in the combination process, let consider the query Q which is associated with four XML elements (e_1, e_2, e_3, e_4) as reported in Table V.4. As can been see, the combined discounting masses for the element e_1 confirms the relevance of this element because each source has ranked e_1 at the first position. However, the element e_2 has been re-ranked (from the fourth rank to the second one) due to its relevance according to the initial information source (S_1) where its score is greater than the score of the element e_3 which is re-ranked at the fourth position.

Element	S_1 *Initial I.R. source*		S_2 *Link I.R. source*		$\alpha_{S_j}(e_i)$		Combined initial masses	Combined discounting masses
	Initial score	Rank	Link score	Rank	*s1*	*s2*		
e_1	0.7	1	0.6	1	1	1	0.778 (1)	0.778 (1)
e_2	0.15	2	0.02	4	0.75	0.25	**0.004 (4)**	**0.089 (3)**
e_3	0.1	3	0.08	3	0.5	0.5	**0.010 (3)**	**0.049 (4)**
e_4	0.05	4	0.3	2	0.25	0.75	0.022 (2)	0.186 (2)

Table V.4 : A simple demonstrative worked example

V.6.3. Fuzzy Formula

The third combination approach we propose is based on fuzzy logic conceptsfor the re-ranking of XML retrieval results by combining both content and link evidences for all retrieved XML elements. Fuzzy logic systems have been attributed with providing an adequate approach for designing robust systems able to deliver an adequate performance when contending with uncertainty. The use of fuzzy logic is principally motivated by the integration of the uncertainty feature of information retrieval.

Based on applying fuzzy processing techniques, some IR approaches have been proposed in the literature. Most of the existing approaches are based on fuzzy set theory, relying on two fuzzy conditions. The first conditions are used for the definition of flexible constraints on queries and stored data (Bratsas, Koutkias, Kaimakamis, Bamidis, & Maglaveras, 2007; Kim & Han, 2009). The second conditions are dedicated for fuzzy ontology with concepts representing the categories and the keywords of a domain (Khokale & Atique, 2013; C.-S. Lee, Wang, & Hagras, 2010). Generally, these approaches do not have adequate ability in uncertainty representation of concepts. In addition, the fuzzy inference rules of these approaches do not consider link evidences. Our fuzzy based approach for the combination of the content and link evidence scores, consider the weighted links based formula for link score computation.

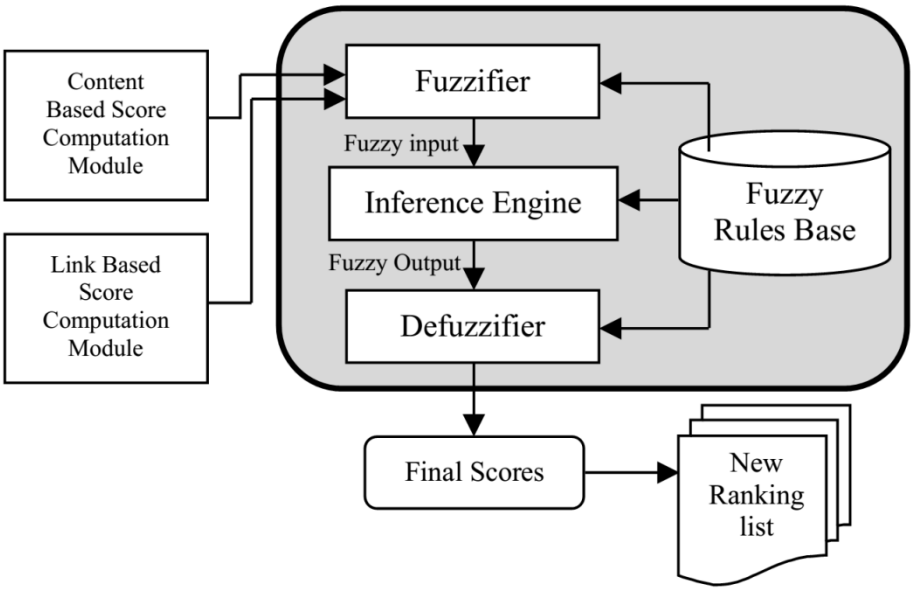


Figure V.6: Score level fusion using content and link evidences.

This section describes the way the combination of two evidence scores (content and link) is done. For facilitating calculations and practical usage, particular fuzzy numbers are used. We consider that the scores computed for both content and link evidences represents an experts'

opinions. The experts' opinions are described by linguistic variables that have been expressed in trapezoidal fuzzy numbers. In our fuzzy system, the antecedents are two variables which are initial and link scores (experts' opinions) of the XML element relevance. We assume that each of these antecedents is represented by three fuzzy sets which are "Low", "Medium" and "High" according to the obtained scores. As shown in Figure V.7, $[S_1, S_2]$ is the interval of thresholds belonging to the fuzzy set "High"; $[S_3, S_4]$ is the interval of thresholds belonging to the fuzzy set "Medium" and $[S_5, S_6]$ is the interval of thresholds belonging to the fuzzy set "low" for both initial and link scores.

The output of the fuzzy system is the XML element relevance possibility which is represented by six fuzzy sets which are "Very Low relevance", "Low relevance", "Medium relevance", "Good relevance", "Very Good relevance" and "Excellent relevance".

Fusion approaches concern every technique for combining outputs of distinct systems (different evidences in our case) and can be accomplished either as a function of retrieval scores, or as a function of the rank in which the retrieved XML element appear. The score-based fusion strategies require normalization among all systems in order to balance the importance of each of them. The used fusion techniques are based on fuzzy logic considering both content and link scores.

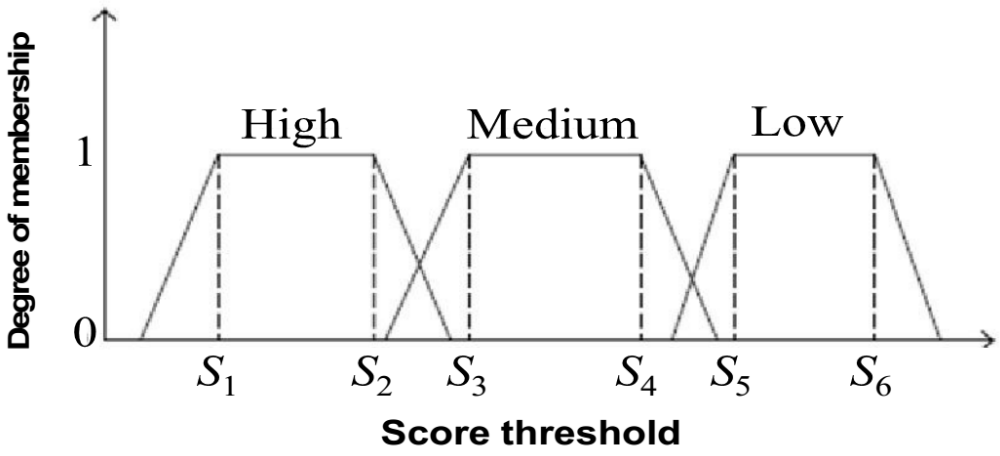


Figure V.7: Fuzzy sets of the proposed entries and their trapezoidal membership functions.

All input variables currently have three fuzzy sets associated with each variable: "High", "Medium" and "Low". The output variable have six fuzzy sets: "Very Low relevance", "Low relevance", "Medium relevance", "Good relevance", "Very Good relevance" and "Excellent relevance".If greater granularity is desired, more fuzzy sets might be defined such as for instance: "very low" and "very high" for each of the two input variables. The proposed and implemented fusion module uses the fuzzy logic principles and methods to combine both content and link scores. The implemented fuzzy inference system is composed of two inputs and one output variables. The output represents the relevance decision taken by our system.

The fuzzy logic conditions for XML element relevance decisions from content and link scores are formulated by a group of nine fuzzy inference rules presented in Table V.5.The order of the rules in fuzzy logic does not affect the output. The fuzzy system uses the knowledge base built with above fuzzy rules. These rules are set according to the following criteria: **(i)** the initial score (*IS*) is more reliable than the link score (*LS*), so we give more importance to the initial score in fuzzy XML element relevance fusion rules; **(ii)** in the cases where the initial score is "Low", the XML element relevance decision should be either "Low" or "Very Low" or

"Medium" even if the link score is "High"; **(iii)** in the cases where the initial score is "High" the document relevance decision should be either "Good" or "Excellent" even if the link score is "Low".

Rules	Input		Output
	Link Score (LS)	Initial Score (IS)	Relevance Decision
R1	Low	Low	Very Low
R2	Low	Middle	Middle
R3	Low	High	Very Good
R4	Middle	Low	Low
R5	Middle	Middle	Good
R6	Middle	High	Very Good
R7	High	Low	Middle
R8	High	Middle	Very Good
R9	High	High	Excellent

Table V.5 : Fuzzy Inference Rules for XML Element Relevance Decisions

In order to explain how the fuzzy inference process is performed each of the steps will be examined:

Input fuzzification. The first step is to take the crisp numerical values of the inputs (*IS* and *LS* scores) and determine the degree to which they belong to each of the appropriate fuzzy sets via the defined membership functions. For instance: content score equal to 0.8would be translated into membership to the fuzzy set "high" and 0.2 would be translated into membership in fuzzy set "low". The implication process input is a single crisp value given by the antecedent. The output is a fuzzy set. It results a fuzzy set represented by a membership function.

In our fuzzy inference system, the mapping between initial and link scores and XML element relevance is accomplished by fuzzy rules. The size of the rule base is 9, constructed via learning from the input/output data by experts. One illustrative fuzzy rule from our rule will have the following structure:

$$R_i: \textbf{\textit{If}} \quad LS \text{ is } Medium \text{ and } IS \text{ is } Low \qquad \textbf{\textit{Then}} \quad Relevance \text{ is } Low$$

Defuzzification. The input for the defuzzification process is the output fuzzy set. Fuzzy set must be defuzzified in order to resolve a single output value from the set. There are various methods for defuzzification such as: smallest of maximum, largest of maximum, the average of the maximum, middle of maximum, centroid and bisector. Among these methods, our approach used centre of area method.

V.7. Conclusions

In this chapter, we have detailed our propositions for the use of link evidence in the XML information retrieval context. These propositions have been divided into 3 parts:

We started our statements by a statistical study that allowed us to conclude that link evidence can play the role of a relevance indicator in the case of Wikipedia XML corpus.

Also, we have focused on the way we calculate the link score of retrieved XML elements. Therefore, we have detailed our three link score computation formulas, which are: adapting Web

link analysis algorithms to the XML context (topical PageRank), distance based approach and weighted links based approach.

Finally, we have described the way in which we introduced the computed link score (or rank) in the computation of the final score of each retrieved XML element. Thus, we have defined three combination methods, which are:linear formula, Dempster-Shafer formula and fuzzy-based formula.

The next chapter will be devoted to the presentation and discussion of the experimental results obtained by our various proposals.

Chapter VI.

Experiments, Results and Discussion

VI.1. Introduction

In this chapter, we present the results of experimental conducted to evaluate our different proposals.

First, we describe the experimental setup and evaluation protocol. The experimental setup will include the details of test collections, tools and evaluation measures. The evaluation protocol will define the way the experimental results have been obtained.

Second, we describe and discuss the results obtained by our different proposals.

Our experiments are organized in three parts: the first part concerns the results obtained by the application of some link analysis algorithms and our "*Topical_Pagerank*" proposition. The second part concerns the results obtained by the "*distance based*" approach with a linear combination formula. The third part concerns the results obtained by the "*weighted links approach*" by using the two proposed combination formula, i.e., "Dempster-Shafer based" and "Fuzzy based".

Many features will be discussed throughout this chapter, for instance:

- The impact of internal links and external links;
- The impact of link score;
- The impact of the quality of initial retrieval results.

VI.2. ExperimentalSetup

We present in this section the experimental setup in which we define the different parameters and experimental conditions used during experimental and evaluation process.

Our experiments were performed using INEX 2007 (indexed using the XFIRM system (Sauvagnat, Hlaoua, & Boughanem, 2006) developed at IRIT (France)) and INEX 2009 Wikipedia XML collections(Denoyer & Gallinari, 2007). The INEX 2007 (INEX 2009 respectively) Wikipedia collection contains 659,388 (2,666,190 respectively) XML documents in English language, densely hyperlinked. Link structure in Wikipedia differs from the Web because it is based on word naturally occurring in a page and link to semantically related page within the collection, i.e., if a word thematic is represented by an article of the collection, word occurrences will automatically link to that article, such as the word "*Italy*" will be a link to the article representing the topic "*Italy*". To evaluate our proposals, we exploit retrieval results (for "Focused" task) from Dalian, Waterloo and MaxPlanck systems related to 107 CAS topics of INEX 2007 (retrieval results of Waterloo University, LIP6 and MaxPlanck institute for 115 CAS topics of INEX 2009) (Geva et al., 2010). This task focuses on the most specific elements, i.e., the user prefers a single element that is relevant to the query even though it may contain some non-specific content, returned as results to the user's query and which are not overlapped. As

recommended in the evaluation in INEX campaign, the metric used in this task is the interpolated Precision at 1% level of recall (iP[0.01]).

VI.2.1. Collection, Data and Tools

VI.2.1.1. Test Collections

a) INEX 2007 Collection

In 2006, Denoyer and Gallinari (Denoyer & Gallinari, 2007) have created a corpus of XML documents based on part of the free encyclopaedia: Wikipedia(Wikipedia). The INEX 2007 Wikipedia XML corpus used at the INEX evaluation initiative contains about659,388 XML documents in English language, densely hyperlinked. This collection is characterized by its linkstructure of a semantic nature wherelinks are based on the occurrence of words in the content of the document.

Figure VI.1 shows an example of XML document extracted from the INEX 2007 Wikipedia collection. It contains many navigational links of type: "*collectionlink*" that point to other XML documents of the INEX Wikipedia collection. The target of a Wikipedia link is indicated by the value of the "*xlink:href*" attribute. On average an article contains 161 XML nodes, where the average depth of a node in the XML tree of the document is 6.72 (Fuhr et al., 2008).

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
       <name id="40774">Base communications</name>
       <conversionwarning>0</conversionwarning>
<body>
       <template name="move to wiktionary"></template>
       <emph3>Base communications </emph3> (basecom):
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple"
       xlink:href="40914.xml"> Communications </collectionlink> services, such as the
       installation, <unknownlink src="operation">operation</unknownlink>,
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple"
       xlink:href="91191.xml">maintenance</collectionlink>, augmentation, modification, and
       rehabilitation of communications networks, systems, facilities, and equipment,
       including off- post extensions, provided for the operation of a military post, camp,
       installation, station, or activity.<emph2>Synonym</emph2>
       <emph3>communications
       <collectionlink xmlns:xlink="http://www.w3.org/1999/xlink"        xlink:type="simple"
xlink:href="510114.xml">base station        </collectionlink>
       </emph3>.
       <p>Source: from <collectionlink      xmlns:xlink="http://www.w3.org/1999/xlink"
       xlink:type="simple" xlink:href="37310.xml">Federal Standard
       1037C</collectionlink>
       </p>
</body>
</article>
```

Figure VI.1: Example of INEX 2007 Wikipedia XML document (file "40774.xml").

b) INEX 2009 Collection

For some of our experiments we carried out by using data from INEX 2009 test collection. This collection comprises 2,666,190 XML documents (a total uncompressed size of 50.7 Gb) and 115 topics (Geva et al., 2010).Starting in 2009, a new document collection based on the Wikipedia has been used. Wikipedia original syntax has been converted into XML format, using both general structural tags ("*article*", "*section*", "*paragraph*", etc.), typographical tags (*emphatic*, *italic*, *bold*, etc.), and frequently occurring link-tags(Geva, Kamps, et al., 2009). The annotation usedhas been enhanced with semantic markup of articles and outgoing links,based on the

semantic knowledge base YAGO, explicitly labelling more than 5,800 classes of entities like "*persons*", "*movies*", "*cities*", etc.The collection contains 101,917,424 XML elements of at least 50 characters (excluding white-space). Figure VI.2 shows an XML document in the INEX 2009 corpus.

```
<?xml version="1.0" encoding="UTF-8"?><!DOCTYPE article SYSTEM "../article.dtd">
<article xmlns:xlink="http://www.w3.org/1999/xlink">
<header>
       <title>Default</title><id>8000</id>
       <revision>
              <id>242931647</id><timestamp>2008-10-04T09:48:59Z</timestamp>
              <contributor><username>Cyfal</username><id>4637213</id></contributor>
       </revision>
       <categories>
              <category>All disambiguation pages</category>
              <category>Disambiguation pages</category>
       </categories>
</header>
<bdy>
<p><b>default</b>,as in failing to meet an obligation, may refer to:
<list>
       <entry          level="1"          type="bullet"><link          xlink:type="simple"
       xlink:href="../gan/Byron_C$enter=2C_M$ichigan.xml">Default (law)</link>
       </entry>
       <entry level="1" type="bullet">
              <link xlink:type="simple" xlink:href="../838/58838.xml">
              Default (finance)</link></entry>
</list>
</p><p><b>default</b>,as a result when no action is taken, may refer to:
<list>
       <entry level="1" type="bullet"><information wordnetid="105816287" confidence="0.8">
              <datum wordnetid="105816622" confidence="0.8"><link xlink:type="simple"
              xlink:href="../316/957316.xml">Default (computer science)</link></datum>
              </information>—also contains consumer electronics usage
       </entry>
       <entry level="1" type="bullet">
              <link       xlink:type="simple"       xlink:href="../639/889639.xml">Default
logic</link>
       </entry>
</list>
</p><p>It may also refer to:
<list>
       <entry level="1" type="bullet">
              <musical_organization wordnetid="108246613" confidence="0.8">
              <group wordnetid="100031264" confidence="0.8">
              <link xlink:type="simple" xlink:href="../344/9159344.xml">Default
       (band)</link></group>
              </musical_organization>,a Canadian post-grunge and alternativerock
       band
       </entry>
       <entry level="1" type="bullet">
              <link xlink:type="simple" xlink:href="../734/3841734.xml">defaults
       (software)</link>,a command line utility for plist (preference) files
       </entry>
</list></p><p>
<table style="background:none">
<row>
       <col style="vertical-align:middle;">
              <imagewidth="30px" src="Disambig_gray.svg"> </image>
       </col>
       <col style="vertical-align:middle;"><it>This  page lists articles associated with
       the        same        title.        If        an<weblink        xlink:type="simple"
       xlink:href="http://localhost:18088/wiki/index.php?title=Special:Whatlinkshere/Defau
       lt&a   mp;namespace=0">internallink</weblink>ledyouhere,  you  may  wish  to  change
       the link to   point directly to the intended article.''</it>
       </col>
</row>
</table></p>
</bdy>
</article>
```

Figure VI.2: Example of INEX 2009 Wikipedia XML document (file "8000.xml").

VI.2.1.2. Topics

The ad hoc topics were created by INEX participants. The created topics contained a short CO query ("*title*" field (keywords)), an optional "*castitle*" field (structure constraints of the topic) (with NEXI syntax), a "*description*" of the search request, and "*narrative*" with a details of the topic of request and the task context in which the information need arose. Figure VI.3 presents an example of an INEX 2007 ad hoc topic. Based on the submitted candidate topics, 107 topics were selected for use in the INEX 2007 collection (topics: 414 to 541) and 115 topics were selected for use in the INEX 2009 collection (topics: 2009001 to 2009115) (Geva et al., 2010).

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_topic topic_id="417"  ct_no="9">
       <title>therapeutic breathing</title>
       <castitle>//*[about(., "therapeutic breathing")]</castitle>
       <description>Find the use of special breathing techniques for therapeutic
       effects</description>
       <narrative>The relevant documents should talk about use of special breathing
       techniques used for therapeutic purpos-     es, either preventive or
curative     ones. These techniques must not use neither drugs nor special devic
       es.Retrieved techniques should include: rebirthing, holotropic    breath
       work,  yoga.</narrative>
</inex_topic>
```

Figure VI.3: Example of INEX 2007 topic format (topic 417).

In Addition to the 2007 topic field, we find a new filled called: "*phrase title*". This field represents a more verbose explanation of the information need given as a series of phrases, just as the title is given as a series of keywords (Geva et al., 2010).

To evaluate our approach we used the returned results related to 107 "Content and Structure" assessed topics of INEX 2007 (topics: 414 to 541) and 115 assessed topics of INEX 2009 (topics: 2009001 to 2009115).

```
<topic id="2009001" ct_no="186">
<title>Nobelprize</title>
<castitle>//article[about(., Nobel prize)]</castitle>
<phrasetitle>"Nobelprize"</phrasetitle>
<description>information about Nobel prize</description>
<narrative>I need to prepare a presentation about the Nobel prize.
       Therefore, I want to  collect information about it as much as possible.
       Information, the history of the     Nobel prize or the stories of the
       award-winners for example, is in demand.
</narrative>
</topic>
```

Figure VI.4: Example of INEX 2009 topic format (topic 2009001).

VI.2.2. Measures

The metric used in the "*Focused task*"of INEX campaign is called interpolated Precision at 1% level of recall (iP[0.01]). This means that the user, in this task, is supposed interested in the most focused results that satisfy its information need from the top XML elements of the retrieved list. Precision is measured as the fraction of retrieved text that was highlighted (by assessors) and recall is measured as the fraction of all highlighted text that has been retrieved.The iP values are computed according to the following formula (Kamps et al., 2008):

$$iP[x] = \begin{cases} \max\limits_{1 \le r \le |L_q|} \left(P[r] \wedge R[r] \ge x \right) & if \ x \le R[|L_q|] \\ 0 & if \ x > R[|L_q|] \end{cases} \quad \text{(VI.1)}$$

Where R[|L_q|] is the recall over all retrieved documents. The P[r] (and R[r] respectively) values represent Precision at rank *r* (Recall at rank *r* respectively) and are computed according to formula II.29 and II.30.

VI.2.3. Pre-Processing

The pre-processing phase in our context consists in preparing all the necessary data for the application of our proposed approaches.

Given that links in INEX 2007 Wikipedia collection are of "*element-document*" link type, because they only point to the roots of documents and not on internal elements, the use of the three link analysis algorithms requires the transformation of "element-document" links to "*document-document*" links. The pre-processing is made to build the graph of links "document-document" that allows us to apply our algorithms "*DOCRANK*" and "*TOPICAL_docrank*". This pre-processing results a new graph containing 659388 nodes, which represent the number of XML documents in the INEX 2007 Wikipedia collection, and 13611471 links of "document-document" type instead of the initial 17 million "element-document" links.

Our "*distance based*" and "*weighted links*" approaches for link score computation are applied in the "*topic-sensitive*" context, which means that they exploit a sub-graph of the global link graph. To obtain this sub-graph, we incorporate two entities, which are: retrieval results and global link graph (*GLG*). We extract firstly all retrieved XML elements (*RXE*) and then we select their respective XML documents (We label this set of documents as: "*A*"). Secondly, we extract all pairs of links between the XML documents "*i*" and "*j*" of "*A*". Next, we build topical link graph "TL*G*" between all XML nodes belonging to "*i*" and "*j*" (which will contain *hierarchical* and *navigational* links). Finally, we add *hierarchical* links between elements of *RXE* belonging to the same XML documents.

VI.3. Evaluation Protocol

To evaluate our resultswe used the INEX evaluation process described in Figure VI.5. As inputs we have:Queries, retrieval results and relevancy assessments for each topic (i.e. *Qrels*). Asoutput we obtained the interpolated precision (*iP* values) at a recall level. The *iP* valueused to compare between systems in the "focused task" of INEX is *iP[0.01]*.

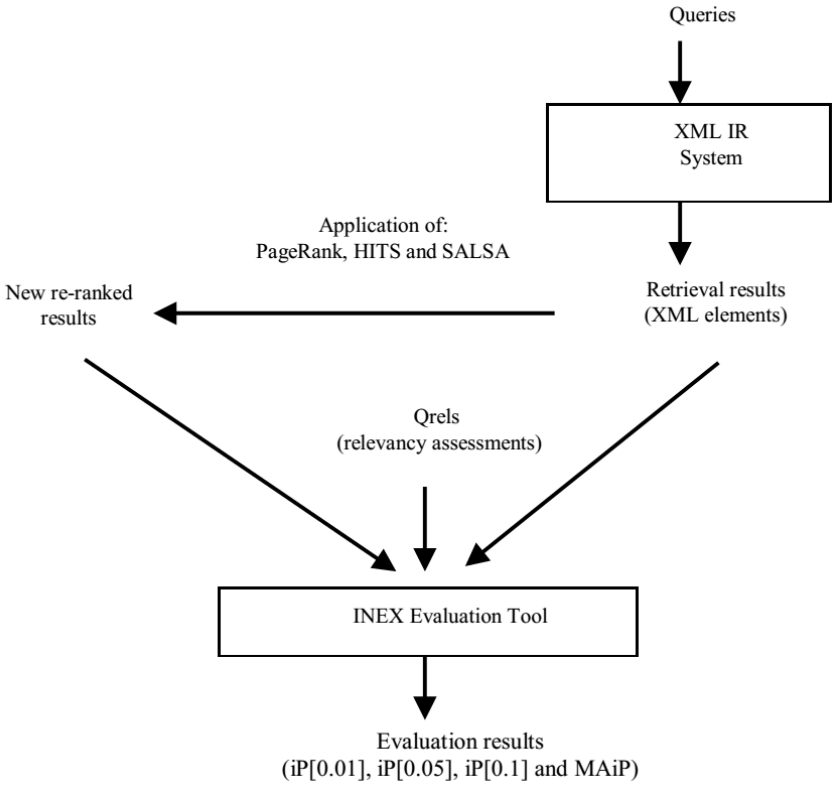


Figure VI.5: INEX Evaluation Process.

For the distance based and weighted links approaches, each experiment is performed following the procedure outlined below:

- Extract the initial retrieval results (from Dalian, Waterloo, MaxPlanck and Justsystem systems of INEX 2007 (INEX); Waterloo, LIP6 and MaxPlanck of INEX 2009 (Geva et al., 2010));
- Construct the topical link graph (internal and external links between retrieved XML elements);
- Compute the link score for each retrieved XML element;
- Normalize the initial and link scores;
- Compute the combined score of each retrieved XML element (the new score);
- Generate the re-ranked list of XML elements;
- Evaluate the new re-ranked list using INEX evaluation tools.

Experiments related to the fuzzy based approach were performed according to the procedure described in the following figure (Figure VI.6).

We have run our approach upon initially returned list, of XML elements, retrieved by the three top ranked systems in the "Focused" task of INEX 2007 (INEX), namely, *Dalian*, *Waterloo* and *MaxPlanck* retrieval systems and a mid-range ranked system: *Justsystem*. For the INEX 2009 data we chose to apply our approach upon the retrieval results of the three best ranked systems of the Focused task, namely, *Waterloo University* (submission: p78-UWatFERBM25F), *LIP6* (submission: p68-I09LIP6Okapi) and the *Max-Planck institute* (submission: p10-MPII-COFoBM) (Kamps, Geva, Trotman, Woodley, & Koolen, 2009).

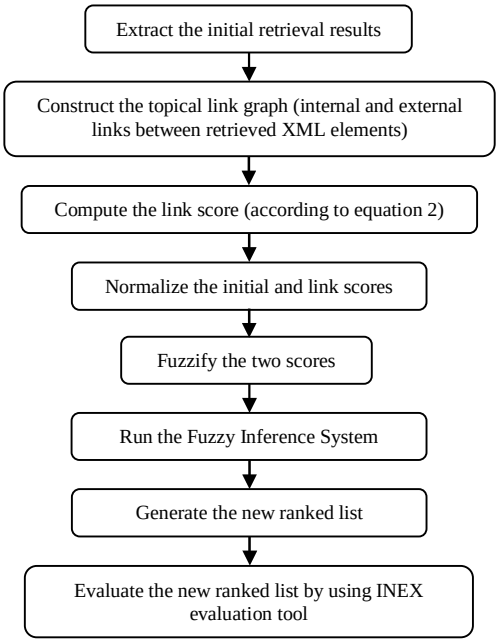


Figure VI.6: Experimental and evaluation process of the fuzzy based approach.

The INEX "Focused" task focuses on the most specific XML elements. An INEX 2007 submission is characterized by three features: the retrieval *task* (focused, relevant in context or best in context); the *topic-id* (specified by the attribute with the same name); the results (composed of: *file*, *path* and *rsv* (or *rank*) tags).

```
<?xml version="1.0"?>
<inex-submission participant-id="26" run-id="DUT_03_Focused" task="Focused" query="automatic"
result-type="element" submission-type="adhoc"><topic-fields title="yes" mmtitle="no"
castitle="yes" description="no" narrative="no"/> <description>cas</description>
<collections><collection>wikipedia</collection></collections>
…
<topic topic-id="424">
       <result>
              <file>27672</file>
              <path>/article[1]/body[1]/section[10]</path>
              <rsv>348.83498706973205</rsv>
       </result>
…
</topic>
…
</inex-submission>
```

Figure VI.7: Example of INEX 2007 retrieval submission format (Dalian retrieval system).

For INEX 2009 the submission format is different. It consists of three types of results: XML elements, file-offset-length (*FOL: File-Offset-Length*) text passages, and ranges of XML elements. The submission format for all tasks is a variant of the familiar *TREC* format extended with two additional fields (*column_7* and *column_8*) (Geva et al., 2010).

topic Q0 file rank rsv run_id column_7 column_8

For the results of type "*Element Results*", a result element may be identified unambiguously using the combination of its file name and the element path in *column_7*. In this case*column_8* will not be used. In the other case, i.e. "*FOL passages*", *column_7* represents the passage offset and *column_8* the length. For "*Ranges of Elements*" result type, the *column_7* represents the start element

path, and *column_8*the end element path. Here are some examples for "*Element Results*", "*FOL passages*" and "*Ranges of Elements*" types of results of INEX 2009 respectively:

1 Q09996 2 0.9998 I09UniXRun1 /article[1]/bdy[1]/sec[2]

1 Q09996 2 0.9998 I09UniXRun1 3892 960

1 Q0 9996 1 0.9999 I09UniXRun1 /article[1]/bdy[1]/sec[1] /article[1]/bdy[1]/sec[3]

VI.4. Experimental Results

VI.4.1. Adapting Web Link Analysis Algorithms: Results and Comments

This section detailed the experimental results obtained by application of three Web link analysis algorithms: Pagerank, HITS and SALSA. In a first step we investigate the use of two variants of the Pagerank Algorithm (global and local) adapted to the XML IR context. Secondly, we show a comparison between these algorithms implemented for XML IR use.

VI.4.1.1. DOCRANK and TOPICAL_docrank results

Our experiments were performed on the results returned by the three of the best ranked retrieval systems in the "Focused" task of INEX 2007, namely, Dalian University of Technology, University of Waterloo and Max-Planck Institüt fur informatik systems. These results are related to the topics: *CO414* to *CO543* (107 "*Content-Only*" topics). The value taken for the *d* parameter is 0.85. The Pagerank algorithm converges generally after 76 iterations with a convergence threshold set to $1E^{-8}$. The execution of the link score computation takes about 30 minutes. The convergence of the "*TOPICAL_docrank*" scores is done in few milliseconds for each topic.

As mentioned, our experiments are conducted on the basis of the results returned by the three systems previously cited for the "Focused" task of INEX 2007. This task focuses on the most specific elements returned as results to the user's query and which are not overlapped. We present the results obtained for theiP[0.01] measurement (which represents the interpolated precision at 1% recall) as recommended at INEX 2007.

The application of "*DOCRANK*" have not showed a significant improvement (0.08% in the best case, with α equal to 0.9 (α is defined in the linear formula, see equationV.13)), which means that the use of links in the global context of the collection does not improve results, but rather the contrary (except when α is equal to 0.9). This statement confirms the results obtained by (Kamps & Koolen, 2008) after applying the "*global indegree*" formula. This is due to the documents that have a high "*DOCRANK*" score and therefore high ranks in all the topics in which they appear as results. An example that we encountered during our experiments is that of the XML document "*31882.xml*" which covers the subject "*United states*". Several topics in which the subject has nothing to do with "*United States*" and for which elements that belong to the "*United States*" document (*31882.xml*) appear in the list of results (because they contain a word of the query), for that case we have seen after application of "*DOCRANK*" that they obtain high scores and consequently higher ranks, which reduces the retrieval accuracy for some topics.So, this phenomenon of infiltration of irrelevant documents causes a reduction in the retrieval quality.

Table VI.1 shows the results obtained after application of "*DOCRANK*" and "*TOPICAL_docrank*" (on several levels depending on the number of documents used in the computation) with variation of α parameter (equation V.13).

DALIAN University DUT_03_Foc	DOCRANK	TOPICAL docrank All returned documents	TOPICAL docrank Top-150 documents	TOPICAL docrank Top-50 documents	TOPICAL docrank Top-20 documents
BaseRun	0.5271	0.5271	0.5271	0.5271	0.5271
α = 0.0	0.4401	0.3424	0.3651	0.3945	**0.4896**
α = 0.1	0,4533	0.3512	0.3920	0.4040	0,4945
α = 0.2	0,4782	0.3725	0.4105	0.4232	0,4959
α = 0.3	0,4939	0.3884	0.4310	0.4420	0,4963
α = 0.4	0,5101	0.4188	0.4540	0.4642	0,4968
α = 0.5	0,5228	0.4383	0.4694	0.4730	0,4974
α = 0.6	0,5256	0.4706	0.4821	0.5102	0,5289
α = 0.7	0,5268	0.4991	0.5185	0.5210	0,5381
α = 0.8	0,5274	0.5221	**0.5305**	0.5343	**0,5470**
α = 0.9	**0,5275**	**0.5300**	0.5296	**0.5356**	0,5351
Best % Improvement	0.08%	0.55%	0.65%	1.61%	3.78%

* *t-test value* = 0.026 = 2.6%

Table VI.1 : iP[0.01] Values obtained after applying "*DOCRANK*" and "*TOPICAL_docrank*" on the results returned by the Dalian University System for several variations of the α parameter (global results for 107 CO topics)

TableVI.1 includes also the results obtained by application of "*TOPICAL_docrank*" (*DOCRANK* in the local context, i.e. query-dependant). These results are better compared to those obtained with "*DOCRANK*" for all values of α and the best rate of improvement is obtained for α equal to 0.8 with the first 20 returned documents for each topic. The best rate of improvement achieved is 3.78%. Improvement of some topics was very significant, for instance: topic 491 and topic 521.

To confirm that these improvements represent a significant rate, we calculated the t-test for all 107 CO topics. The p-value obtained is equal to 2.6%, confirming that the improvement is significant even if relatively low.

The next figure show the "TOPICAL_docrank" results (compared withbaseline) obtained for some CO topics with α=0.8 by using the top 20 retrieved documents (retrieval results of Dalian XML IR System). Most of these topics has been improved by the "TOPICAL_docrank" approach.

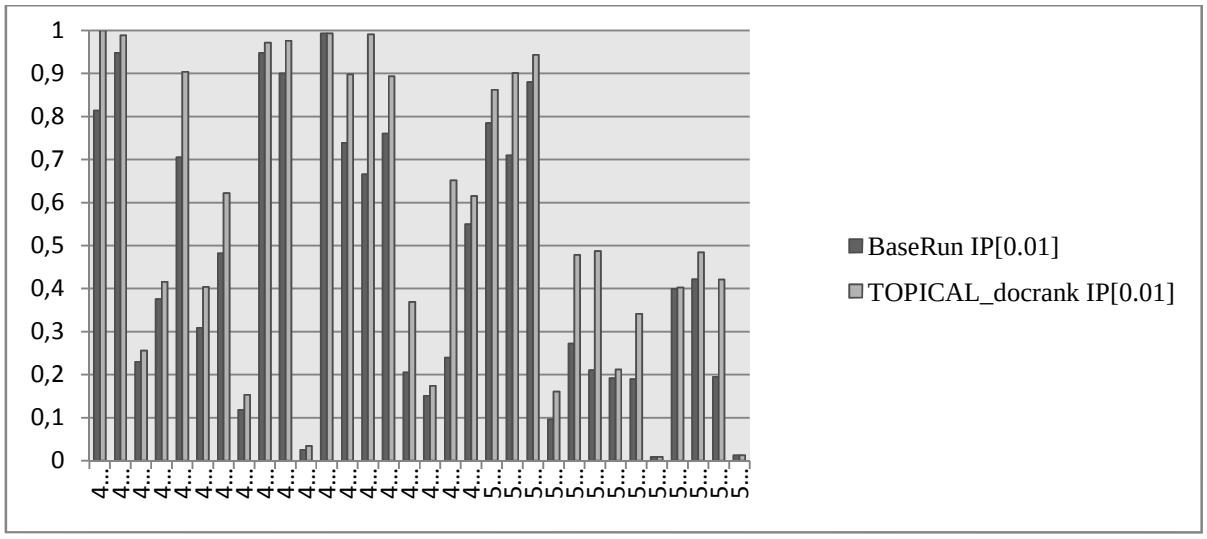


Figure VI.8: Baseline and "*TOPICAL_docrank*" results obtained for some CO topics with α=0.8 and number of documents=20 (Dalian University System)

To approve the improvements achieved by application of "*TOPICAL_docrank*" on the results returned by the DALIAN XML IR system, we have applied this algorithms to two other systems ranked among the top systems in INEX 2007 for the "Focused" task.

WATERLOO University	DOCRANK	TOPICAL docrank Top-150 documents	TOPICAL docrank Top-50 documents	TOPICAL docrank Top-20 documents
BaseRun	0.5108	0.5108	0.5108	0.5108
α = 0.0	0.3710	0.3651	0.3945	0.4816
α = 0.1	0,3940	0.3940	0.4425	0,4899
α = 0.2	0,4056	0.4052	0.4482	0,4912
α = 0.3	0,4142	0.4115	0.4523	0,4931
α = 0.4	0,4385	0.4378	0.4691	0,4946
α = 0.5	0,4534	0.4602	0.4714	0,4953
α = 0.6	0,4698	0.4770	0.4787	0,5073
α = 0.7	0,4808	0.4890	0.4910	0,5179
α = 0.8	0,4992	0.4992	0.4948	**0,5218**
α = 0.9	0,5100	0.5100	**0.5135**	0,5001
Best % Improvement	-0.001%	-0.16%	0.53%	2.15%

Table VI.2 : iP[0.01] Values obtained after applying "*TOPICAL_docrank*" on the results returned by the University of Waterloo System

Table VI.2 shows the iP[0.01] values obtained after applying "*TOPICAL_docrank*" on the results returned by University of Waterloo retrieval system. These results confirm those obtained with DALIAN retrieval system, and the best rate of improvement is obtained with the same parameters as the first system (α = 0.8 and number of documents=20 XML documents).

Table VI.3 represents the iP[0.01] values obtained after applying "*TOPICAL_docrank*" on the results returned by the Max Planck Institute retrieval system.

MAX-PLANCK Institut fur informatik	DOCRANK	TOPICAL docrank Top-150 documents	TOPICAL docrank Top-50 documents	TOPICAL docrank Top-20 documents
BaseRun	0.5066	0.5066	0.5066	0.5066
α = 0.0	0.3634	0.3705	0.4180	0.4515
α = 0.1	0,3775	0.3775	0.4366	0,4646
α = 0.2	0,3894	0.3914	0.4512	0,4683
α = 0.3	0,4025	0.4221	0.4546	0,4706
α = 0.4	0,4211	0.4502	0.4594	0,4798
α = 0.5	0,4356	0.4633	0.4641	0,4874
α = 0.6	0,4514	0.4704	0.4702	0,4902
α = 0.7	0,4708	0.4792	0.4760	0,4926
α = 0.8	0,4822	0.4822	0.4792	0,4954
α = 0.9	0,5000	0.5000	0.5027	**0,5072**
Best Improvement %	-0.01%	-1.30%	-0.77%	0.12%

Table VI.3 : iP[0.01] Values obtained after applying "*TOPICAL_docrank*" on the results returned by the MAX-PLANCK Institut fur informatik System

The rate of improvement is less significant compared to the rates obtained with the two first systems. This is due to the retrieval strategy adopted by the system of the MaxPlanck Institute. This strategy is based on "*CAS-title*"[16] information of topics. This eliminates many documents from the list of *top-N* elements returned because they do not meet the structural constraints mentioned in the "*CAS-title*" of topics. We found here that in most of topics there are no links between the *top-N* documents returned, which justifies the retrieval quality after application of "*TOPICAL_docrank*".

From the three previous tables of evaluation results we have seen that the increase in the value of the α parameter (in other words the impact of "*DOCRANK*" and "*TOPICAL_docrank*" scores are reduced) makes the retrieval quality better, which means that the textual information (the scores initially assigned to the XML elements by the retrieval system) is more important compared to XML links information. Best experimental improvements are obtained with α=0.8 (see equation V.13).

For all the three systems the best rate of improvement is obtained by using the top 20 documents returned for each topic. This may be due to the reduction of the phenomenon of infiltration of irrelevant documents that we have already mentioned (in other words, if a document is pointed throughout the collection with 1000 links, it will be pointed at most by 19 documents in the set of 20 first documents returned by the retrieval system and these documents are considered as the best for the topic in question).

[16]CAS-Title: Content-And-Structure Title, this topic information indicates the structural constraints of the topic (query)

VI.4.1.2. Comparison between PageRank, Topical_PageRank, HITS and SALSA

In the followingsection we present a comparative evaluation between the re-ranked lists of XML elements obtained by application of the three link analysis algorithms, and the initially returned lists by the Dalian, Waterloo and MaxPlanck systems.

In tablesVI.4, VI.5 and VI.6, we present iP[0.01] values obtained for the baseline and by application of the three link analysis algorithms. The columnsrepresent the baseline and the variation of the α parameter (equation V.13). The rows represent the link analysis algorithms used in these experiments.The column "*Topical_Pagerank*" of the next three tables concerns the evaluation results obtained by application of PageRank in the topical context, i.e., on subset of the returned list of XML elements. Generally, we use the top 20 returned documents. This algorithm shows improvements for all retrieval systems. The best improvement was about 3.78% obtained with α=0.8. The rate of improvement for the MaxPlanck retrieval system is less significant with "*Topical_Pagerank*" compared to Dalian or Waterloo systems.

Run	PageRank	Topical_PageRank	HITS	SALSA
Baseline	0.5271	0.5271	0.5271	0.5271
α=0.0	0.4401	0.4896	0.3217	0.4910
α=0.1	0.4533	0,4945	0.3545	0.5033
α=0.2	0,4782	0,4959	0,3828	0,5064
α=0.3	0,4939	0,4963	0,4011	0,5100
α=0.4	0,5101	0,4968	0,4165	0,5126
α=0.5	0.4828	0,4974	0.4464	0.5188
α=0.6	0,5256	0,5289	0,4815	0,5207
α=0.7	0,5268	0,5381	0,5123	0,5286
α=0.8	0.5274	**0,5470***	**0.5366**	**0.5478****
α=0.9	**0.5275**	0,5351	0.5273	0.5346

* : t-test value = 2,6 %
** : t-test value ≅ 5 %

Table VI.4 : iP[0.01] Values obtained by baseline and by application of the link analysis algorithms on results returned by the Dalian system

Run	PageRank	Topical_PageRank	HITS	SALSA
Baseline	0.5108	0.5108	0.5108	0.5108
α=0.0	0.3710	0.4816	0.2721	0.4682
α=0.1	0.394	0,4899	0.2760	0.4797
α=0.8	0.4992	**0,5218**	0.5054	**0.5513**
α=0.9	0.5100	0,5001	**0.5117**	0.5283

Table VI.5 : iP[0.01] Values obtained by baseline and by application of the link analysis algorithms on results returned by the Waterloo system

Run	PageRank	Topical_PageRank	HITS	SALSA
Baseline	0.5066	0.5066	0.5066	0.5066
α=0.0	0.3634	0.4515	0.2813	0.4405
α=0.1	0.3775	0,4646	0.2963	0.4463
α=0.8	0.4822	0,4954	0.4889	**0.5278**
α=0.9	0.5000	**0,5072**	0.4959	0.5171

Table VI.6 : iP[0.01] Values obtained by baseline and by application of the link analysis algorithms on results returned by the MaxPlanck system

As we see in the HITS column of tables VI.4, VI.5 and VI.6, improvement was less important for the three systems. The best rate of improvement was about 1.8% for Dalian system with α=0.8. This may be due to the TKC[17] (*Tightly Knit Community*) effect which in certain cases prevents the HITS algorithm from identifying meaningful authorities (Lempel & Moran, 2000)(Lempel & Moran, 2001). Application of HITS on MaxPlanck retrieval system results do not show improvement, this may be due to the same previously mentioned reason (i.e. *CAS-Title* strategy).

The last column of tables VI.4, VI.5 and VI.6 represents iP[0.01] values obtained by application of the *SALSA* algorithm. The results show that *SALSA* performs better than the other algorithms (*PageRank*, *Topical_pagerank* and *HITS*) in the most of cases. This may be due to the combination of random walk principle (defined in the PageRank algorithm) and the hub and authorities principle (defined in *HITS*). *SALSA* resolves also the *TKC* effect. The best improvements were about: 3.92% for the Dalian retrieval system, 7.92% for the Waterloo system and 4.18% for the MaxPlanck system. These rates are all obtained for α = 0.8.

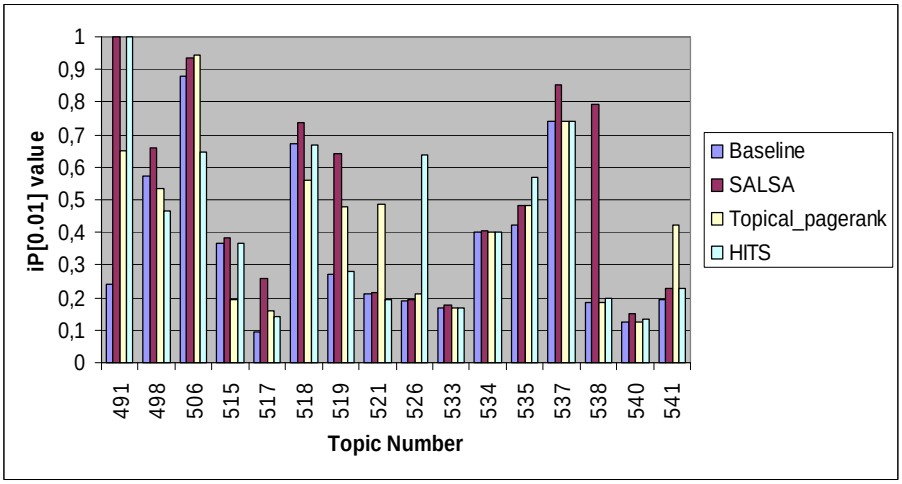


Figure VI.9: iP[0.01] values obtained for some Content Only (CO) topics by application of the link analysis algorithms on DALIAN system results, with α = 0.8.

[17] TKC effect is a small but highly inter connected set of pages. It occurs when such a community scores high in link analysis algorithms, even though the pages in the TKC are not authoritative on the topic or pertain to just one aspect of the topic.

We have calculated the t-test, i.e., for 107 topics to check the significance of the improvements. The obtained t-test value for the *Topical_pagerank* was equal to 2.6% and 5% for *SALSA*. These t-test values confirm that the improvements are significant.

If we analyse the evaluation results topic by topic we conclude that *SALSA* performs better in the most of topics (more than 50% of the 78 improved topics) compared to *PageRank*, *Topical_PageRank* and *HITS (see* Figure VI.9*)* .

To summarize, we can say that: links are a source of evidence which plays an important role in determining the relevant elements in the XML IR context. Also, we found that the "topic-sensitive" approaches give better results compared to global approaches. We studied the impact of variation of α parameter, and we confirmed that the combination of the two scores, i.e., initial score and link score, performsbetter than each taken separately.

In the next sections, we will detailed the results obtained by our two "topic-sensitive" approaches exploiting "element-element" links.

VI.4.2. Distance Based Approach: Results and Comments

In this section we describe and discuss the experimental results obtained by the "*distance based*" approach. We particularly describe, the impact of internal links, the impact of link score and the impact of the initial retrieval list quality.

VI.4.2.1. Impact of internal links

As defined in equationsV.5 and V.6, our approach includes both internal (hierarchical) links and navigational links in the computation of each XML element link score. In this experiment we attempt to understand the impact of these links. Whether it is interesting to consider only internal links emerging from XML elements receiving external links or all internal links between the retrieved XML elements?Also, we attempt to identify the best value of β parameter of equation V.5.

For this series of experiments we set α parameter to 0,6 (this value represents the best value of α obtained by experiments described in the next section, i.e. impact of link score) and we used the initial results retrieved by Dalian system of INEX 2007.

To define the best way to consider internal links, we evaluated two variants of inclusion of these links: **(a)** include all internal links between the retrieved XML elements; **(b)** include only internal links emerging from XML elements receiving external links. To identify the best variant, we run our approach upon the 107 topics for both variants. Table VI.7 lists iP[0.01] values obtained for both variants.

Baseline	variant (a)	variant (b)
0,5271	0,5434	0,5715 (+8,42%)

Table VI.7 : iP[0.01] values obtained for both internal links inclusion alternatives (with: α=0,6; β=0,2; INEX 2007)

Results confirm that the best variant is the second (*variant (b)*). We found that multiple XML elements have obtained high scores because they belong to documents containing a lot

ofelements returned as retrieval results. For instance in topic 423, eleven (11) XML elements belonging to the XML document "*12207.xml*" (assessed as not relevant for this topic) are returned as retrieval results. Using the first variant *(a)*, the propagated relevance score between these XML elements will be high, and consequently, the computed link scores will be high, which resulted in lower accuracy for that topic compared to variant *(b)*. The reason here is simple: if there is no external inlinks, this means that theXML element is unlikely to be relevant. The variant *(b)*will be used for all coming experiments.

For the second feature (*β parameter*), we varied its value from 0 to 1. We found that the best improvement is obtained when β is equal to 0,2 (Table VI.8). This means that the external links are more important, in our context, than the internal links. But the combined score of the two link types is more effective compared to the use of only external links.

We notice from Table VI.8 that the use of internal links (with most values of β parameter) improves the retrieval accuracy compared to the case without using internal link information.

β	**iP[0.01] over all (107) topics**
(external links only) β=0,0	0,5430
β=0,1	0,5627
β=0,2	**0,5715 (+8,42%)***
β=0,3	0,5621
β=0,4	0,5522
β=0,5	0,5510
β=0,6	0,5469
β=0,7	0,5391
β=0,8	0,5331
β=0,9	0,5310
- Paired sample T-test: * Significant (p<0.05) against baseline	

Table VI.8 : iP[0.01] values obtained by variation of β parameter (Dalian system retrieval results of INEX 2007 with α=0,6)

VI.4.2.2. Impact of link score in the final score computation formula

Our main aim in this section is to experiment the best value of α parameter that determines the degree of contribution of each score (initial or link scores) in the final score (equation V.13). In all the following tables β value is fixed to 0.2 (see Table VI.8).

In this section we compare the results obtained by our approach with those obtained when applying some link analysis algorithms we implemented and adapted to XML links (M'hamed Mataoui et al., 2010) (see section V.3). The "*Topical_Pagerank*" (respectively "Pagerank") columns represent related results obtained by our implementation of the Pagerank algorithm (Brin & Page, 1998)in the topical context (respectively global context). HITS(Kleinberg, 1999) and SALSA (Lempel & Moran, 2001) columns represent the results obtained by our implementation of these two algorithms. All these algorithms exploit "*document-document*" link type contrary to our "*distance based*" approach that exploits "*element-element*" link type.

The most important finding from table VI.9 is that our approach outperforms the other approaches with all the variations of the α parameter. The paired-sample *t-test* computed between the results of our approach and the baseline over the 107 topics (each of the topics is treated

separately) show that our distance based approach is statistically significant compared to the baseline and other algorithms.

Run	Baseline	Pagerank	Topical Pagerank	HITS	SALSA	Distance Based
α = 0.0		0.4401	0.4896	0.3217	0.4910	0,4953
α = 0.1		0,4533	0,4945	0,3545	0,5033	0,5274
α = 0.2		0,4782	0,4959	0,3828	0,5064	0,5304
α = 0.3		0,4939	0,4963	0,4011	0,5100	0,5338
α = 0.4		0,5101	0,4968	0,4165	0,5126	0,5341
α = 0.5	0,5271	0,5228	0,4974	0,4464	0,5188	0,5362
α = 0.6		0,5256	0,5289	0,4815	0,5207	**0,5715(+8,42%)** *,¥
α = 0.7		0,5268	0,5381	0,5123	0,5286	0,5575
α = 0.8		0,5274	**0,5470 (+3,78%)**	**0,5366 (+1,8%)**	**0,5478 (+3,92%)**	0,5428
α = 0.9		**0,5275 (+0,08%)**	0,5351	0,5273	0,5346	0,5390

- values between parenthesis represent the percentage of accuracy improvement compared to baseline
- Paired sample t-test: * Significant ($p<0.05$) against baseline, Pagerank, Topical_pagerank, HITS and SALSA
- ¥ : Baseline MAiP = **0,1689** ; Distance based MAiP = **0,2117** (**+25,34%**)

Table VI.9 : iP[0.01] values obtained by distance based approach compared to the other link analysis algorithms (application on Dalian system retrieval results of INEX 2007)

In the case of Dalian retrieval results, we obtained improvements over more than 72 topics, which represent 67% of the 107 topics (Table VI.10). Compared to "Topical_ Pagerank" algorithm that has improved only 45 topics, our approach is better. Only 12 topics were not improved, some of them are listed in Table VI.11. The best rate of improvement is obtained with α=0,6 and β=0,2.

Discussion about The two variants of Pagerank, HITS and SALSA can be found in section VI.4.1.

In this context, i.e., studying the link score impact, we carried out experiments about the impact of using combination formula (the same linear formula, i.e., equationV.12) based on "element rank" information instead of the scores. The obtained re-ranking results did not show any improvement of the retrieval accuracy but rather the contrary (diminution about -10%). We believe that part of the obtained improvement is due to the good choice of the combination formula as well as the used parameters.

In the following tables (Table VI.10 and Table VI.11), we present respectively the best improvements and the worst diminution on iP[0.01] values of some INEX 2007 topics. The improvements are very significant for several topics, like: 441, 450, 459, 489, 491, 496, 515, 517, 522 and 527. However, we found that there are some topics that have not been improved. Some of these topics are listed in Table VI.11.

Most of the decreases in relevance are caused by the phenomenon of absence of external links between the XML elements returned as retrieval results for some topics. As an example, we can cite the case of topic 414, in which there was only two navigational links between all the retrieved XML elements. This means that only three XML elements will have a good link score and the rest of XML elements will have only the minimum score according to the first part of the equation V.5.

Topic number	Baseline	Topical Pagerank	Distance Based (Improvement %)
Topic 418	0,7197	0,7197	0,9188 (+27,66%)
Topic 419	0,6391	0,6391	0,753 (+17,82%)
Topic 425	0,8141	1	0,892 (+9,56%)
Topic 441	0,2297	0,2559	0,7466 **(+225,03%)**
Topic 448	0,7052	0,9038	0,94 (+33,29%)
Topic 450	0,4644	0,4462	0,9897 **(+113,11%)**
Topic 454	0,309	0,4044	0,4731 (+53,10%)
Topic 458	0,4819	0,6224	0,5783 (+20,00%)
Topic 459	0,3485	0,2968	0,9183 **(+163,50%)**
Topic 473	0,1181	0,1532	0,1871 **(+58,42%)**
Topic 480	0,4679	0,3239	0,5666 (+21,09%)
Topic 481	0,0252	0,0341	0,04272 **(+69,52%)**
Topic 484	0,6659	0,9914	0,9914 (+48,88%)
Topic 488	0,7605	0,8939	1 (+31,49%)
Topic 489	0,2057	0,3693	0,9984 **(+385,36%)**
Topic 490	0,151	0,1744	0,2393 **(+58,47%)**
Topic 491	0,2398	0,652	0,5857 **(+144,24%)**
Topic 496	0,5497	0,6154	1 **(+81,91%)**
Topic 515	0,3684	0,1947	0,5602 **(+52,06%)**
Topic 517	0,0959	0,1608	0,2869 **(+199,16%)**
Topic 521	0,2107	0,4873	0,5816 **(+176,03%)**
Topic 522	0,5564	0,5564	1 **(+79,72%)**
Topic 527	0,1897	0,3413	0,3387 **(+78,54%)**
Topic 534	0,3999	0,4028	0,4982 (+24,58%)
Topic 535	0,4221	0,4842	0,5937 (+40,65)
Topic 537	0,7419	0,7419	0,8677 (+16,95%)
Topic 541	0,1953	0,4208	0,2785 (+42,60%)
Topic 542	0,0026	0,0023	0,00338 (+30%)

Table VI.10 : Best topic by topic improvement iP[0.01] Values obtained by application of our distance based approach, compared to baseline and Topical_ Pagerank results (application on Dalian system retrieval results)

Topic number	Baseline	Distance Based (Diminution %)
Topic 414	1	0,4204 (-57,96)
Topic 416	0,0469	0,0453 (-3,32%)
Topic 471	1	0,1510 (-84,9%)
Topic 487	0,0792	0,0681 (-14,01%)
Topic 498	0,5728	0,5275 (-7,90%)
Topic 516	0,8429	0,6085 (-27,80%)
Topic 518	0,6731	0,5605 (-16,72%)

Table VI.11 : Some iP[0.01] Values (diminution) obtained by application of our distance based approach, compared to baseline (application on Dalian system retrieval results of INEX 2007)

VI.4.2.3. Impact of the initial retrieval results

We experimented also our "*distance based*" approach by considering other initial retrieval results, namely those returned by the Waterloo university retrieval system (ranked 3rd in INEX 2007) and MaxPlanck institute retrieval system (ranked 5th in INEX 2007).

Table VI.12 shows the iP[0.01] values obtained by our approach on the retrieval results returned by the Waterloo university retrieval system. We notice that our approach performs

better in this case compared to the other implemented algorithms and gets the best rate of improvement with the configuration: α=0,6 (for which the improvement was about +7,98%).

Table VI.13 shows the iP[0.01] values obtained by our approach on the retrieval results returned by the MaxPlanck institute system. We have also obtained in the case of MaxPlanck institute retrieval system the best improvements over the other algorithms. The best improvement rate was about +5,42% (with α=0.6).

As we noted in (M'hamed Mataoui et al., 2010), the results returned by the MaxPlanck system are related to structural constraints specified in the tag "<*castitle*>" of each topic. According to the description documents of INEX topics, there are a lot of topics that require returned elements of type (tag) "*article*". In this case the application of our approach is equivalent to "*Topical_Pagerank*" approach because "*article*" elements will point only "*article*" elements.

Run	Baseline	Pagerank	Topical Pagerank	HITS	SALSA	Distance Based
α = 0.0		0.3710	0.4816	0.2721	0.4682	0,4906
α = 0.1		0,3940	0,4899	0,2760	0,4797	0,5144
α = 0.2		0,4056	0,4912	0,2879	0,4881	0,5174
α = 0.3		0,4142	0,4931	0,3006	0,4975	0,5213
α = 0.4		0,4385	0,4946	0,3344	0,5067	0,5245
α = 0.5	0,5108	0,4534	0,4953	0,3642	0,5134	0,5267
α = 0.6		0,4698	0,5073	0,4371	0,5211	**0,5516 (+7,98%)*,**,¥**
α = 0.7		0,4808	0,5179	0,4512	0,5341	0,5331
α = 0.8		0,4992	**0,5218 (+2,15%)**	0,5054	**0,5513 (+7,92%)**	0,5247
α = 0.9		0,5100	0,5001	**0,5117 (+0,2%)**	0,5283	0,5178

- Paired sample T-test: * p<0.05 against baseline, Pagerank, HITS, Topical_pagerank
** p<0.5 against SALSA
- ¥ : Baseline MAiP = **0,1764** ; Distance based MAiP = **0,2109** (**+19,55%**)

Table VI.12 : iP[0.01] values obtained by distance based approach compared to the other link analysis algorithms (application on Waterloo system retrieval results of INEX 2007)

Run	Baseline	Pagerank	Topical Pagerank	HITS	SALSA	Distance Based
α = 0.0		0.3634	0.4515	0.2813	0.4405	0,4912
α = 0.1		0,3775	0,4646	0,2963	0,4463	0.5047
α = 0.2		0,3894	0,4683	0,3216	0,4572	0,5055
α = 0.3		0,4025	0,4706	0,3481	0,4706	0,5079
α = 0.4		0,4211	0,4798	0,3646	0,4843	0,5106
α = 0.5	0,5066	0,4356	0,4874	0,3800	0,4952	0,5251
α = 0.6		0,4514	0,4902	0,4421	0,5084	**0,5341 (+5,42%)*,**,¥**
α = 0.7		0,4708	0,4926	0,4673	0,5124	0,5293
α = 0.8		0,4822	0,4954	0,4889	**0.5278 (+4,18)**	0,5287
α = 0.9		0,5000	**0,5072 (+0,11%)**	0,4959 (-0,02%)	0,5171	0,5247

- Paired sample T-test: * p<0.01 against baseline, pagerank, topical_pagerank and HITS
** p<0.05 against SALSA
- ¥ : Baseline MAiP = **0,1307** ; Distance based MAiP = **0,1678** (**+28,38%**)

Table VI.13 : iP[0.01] values obtained by distance based approach compared to the other link analysis algorithms (application on MaxPlanck system retrieval results of INEX 2007)

We conducted another test to evaluate the behaviour of our approach upon the retrieval results returned by a mid-range ranked system of the INEX 2007 initiative. We have chosen for this test the retrieval results returned by the 10th ranked system: Justsystem retrieval system. We used the best configuration parameters, i.e., α=0,6 and β=0,2. The following table shows the obtained results.

Baseline	Distance Based
0,4631	0,5014 **(+8,27%)**

- Paired sample T-test: * p<0.05
- ¥ : Baseline MAiP = **0,1196** ; Distance based MAiP = **0,1353** (**+13,12%**)

Table VI.14 : iP[0.01] values obtained by distance based approach compared to the baseline (application on JustSystem retrieval results of INEX 2007)

As we see in Table VI.14, evaluation results confirm that, even if we apply our approach on the retrieval results returned by a mid-range ranked system, the improvements are significant compared to baseline.

VI.4.2.4. Evaluation of the Robustness of our Approach

As our approach need to set some parameters, namely, α and β, we conductedthe 5-fold cross-validation test applied upon the Dalian system retrieval results; the second is the application of our approach upon the retrieval results of INEX 2009 test collection(INEX).

To do the first validation test we partitioned the topics of our INEX 2007 collection into five subsets. These five subsets will be used to compose two sets. The first set is called "*Training set*" and the other set called "*Test set*". The aim of this cross-validation test is to find the best configuration (α and β values) in the training set and apply this configuration to the test set to observe the behaviour of our approach. The five subsets are:

- Subset 1: 27 topics (topic 414 to topic 445)
- Subset 2: 20 topics (topic 446 to topic 473)
- Subset 3: 20 topics (topic 474 to topic 496)
- Subset 4: 20 topics (topic 497 to topic 522)
- Subset 5: 20 topics (topic 523 to topic 543)

A single subset is retained as the test set for testing the approach, and the 4 remaining subsets are used as training set. The best configuration in the training test for all folds was α=0.6 and β=0.2. With this configuration (α=0.6 and β=0.2) applied to the test sets, we obtain iP[0.01] value equal to: 0,5676 (+9,44% of accuracy improvement). This means that our α and β parameters are robust to the collection change and improve retrieval accuracy in all the experimented cases.

Training Set		Test Set	
Baseline	**Distance Based Approach**	**Baseline**	**Distance Based Approach**
0,5264	0,5734 (+8,92%)*	0,5186	0,5676 (+9,44%)**

- Paired sample T-test:
  * p<0.01 ** p<0.1

Table VI.15 : iP[0.01] values obtained by distance based approach over the training and test sets (application on Dalian system retrieval results of INEX 2007)

The second series of validation test is carried out by applying our approach on another test collection, namely, INEX 2009. This collection comprises 2,666,190 XML documents and 115 topics (Kamps & Koolen, 2008). We chose to apply our approach upon the retrieval results of the three best ranked systems of the "*Focused task*", namely, Waterloo University (submission: p78-UWatFERBM25F), LIP6 (submission: p68-I09LIP6Okapi) and the Max-Planck institute (submission: p10-MPII-COFoBM). We also used in this validation test the best configuration parameters obtained for the INEX 2007 collection, i.e., α=0,6 and β=0,2 to evaluate our approach upon the INEX 2009 collection retrieval results. The following table shows the obtained results.

We mentioned in Table VI.16 the iP[0.01] and MAiP values, obtained by the baseline and our approach, for each of the three systems. We notice improvements for iP[0.01] measurement which varies from 2,58% up to 4,41%. Concerning the rates of improvement obtained by MAiP measurement, the values show that they are more significant (about 16%). All that evaluation results shows that our approach allows, by applying the same parameters (α and β) obtained for INEX 2007 collection, to improve accuracy over INEX 2009 retrieval results.

Run	**Waterloo Baseline**	**Distance Based**	**LIP6 Baseline**	**Distance Based**	**Max-Planck Baseline**	**Distance Based**
iP[0.01]	0,6333	0,6497 (+2,58%)	0,6141	0,6412 (**+4,41%**)*	0,6134	0,6348 (+3,48%)
MAiP	0,1854	0,2151 (**+16,01%**)**	0,3001	0,3152 (+5,03%)	0,1973	0,2283 (+15,71%)

- Paired sample T-test: *,** p<0.1

Table VI.16 : iP[0.01] and MAiP values obtained by distance based approach (application on the three best ranked systems in the focused task of INEX 2009 with α=0,6; β=0,2)

VI.4.3. Weighted Links Based Approach: Results and Comments

VI.4.3.1. Dempster-Shafer Combination Formula

To evaluate our proposal we carried out a set of experiments based on INEX test collection.We present some results obtained by our approach compared to baseline as shown in Table VI.17 and Table VI.18.

From Table VI.17, we note that "*Weighted links+Dempster-Shafer formula*" proposed approach improves accuracy in most of the topics (i.e. 416, 419, 421, 422, 425, 473, etc.). Thanks to the Demspter–Shafer theory and the link computation approach, the obtained combined results show significant improvement compared to baseline, which conclude to that link evidence plays an important role as an accurate source of evidence in the XML elements relevance computation process.

Topic Id	Baseline	Combined mass (DS)	Improvement %	Combined mass (DS) with discounting rate	Improvement %
414	1	0,4204	-57,96	0,4204	-57,96
415	0,5525	0,2333	-57,77	0,2094	-62,10
416	0,0469	0,07258	54,75	0,05871	25,18
417	0,0005	0,0005	0	0,0005	0
419	0,6391	0,7104	11,16	1	56,47
421	0,4175	1	139,52	0,634	51,85
422	0,0386	0,0533	38,08	0,03867	0,18
424	1	1	0	1	0
425	0,8141	1	22,83	1	22,83
426	0,8372	1	19,44	1	19,44
428	1	1	0	1	0
429	0,9479	1	5,49	1	5,49
433	1	0,6188	-38,12	0,7138	-28,62
434	0,9798	0,9812	0,14	0,9812	0,14
436	0,0173	0,0173	0	0,0173	0
473	0,1181	0,1459	23,53	0,1435	21,50
521	0,2107	0,4873	131,27	0,3128	48,45

Table VI.17 : iP[0.01] values & improvement obtained by application of the combined DS theory (Dalian system retrieval results, some topics)

We observe that some topics in which improvement is equal to 0 is principally due to the value of the baseline. Our approach gives the same highest value (*iP[0.01] = 1*), and as a consequence, it confirms the importance of the value of the content evidence. In this case, the link evidence supports the content evidence. However, in the case of topics 417 and 436, the non-improvement is due to the lowest accuracy of the initially retrieved results (topic 417: iP[0.01] =0.0005). Most of the relevance decreases in Table VI.17 are due to the absence of navigational links between returned XML elements. For instance, topics 414 and 433 which have a baseline iP[0.01] equal to 1, contain only two navigational links. This means that link evidence cannot contribute in the selection of relevant elements, because only few elements will get a high link score.

Baseline	Topical Pagerank	Combined mass (DS)	Combined mass (DS) with discounting rate
0.5271	0,5470 (+3.78%)	0.5682 (+7.79%)	0.5591 (+6.07%)

Table VI.18 : iP[0.01] values obtained by combined DS theory compared to baseline and Topical Pagerank (Dalian retrieval resultsover all topics)

According to Table VI.17 and Table VI.18, we note that the two variants of combination (with and without discounting rate) improve the retrieval accuracy, and the variant without the discounting rate outperforms the one using the discounting except for some topics (419, 433, etc.).

Compared to "*Topical Pagerank*" approach (Mataoui et al., 2010), the two combination DS variants performs better. These results can be interpreted by the use of the "*element-element*" link type instead of "*document-document*" link type (used by *"Topical Pagerank"*).

A multitude of discounting rate formulas can be proposed in order to define an appropriate value allowing best improvements.

VI.4.3.2. Fuzzy Combination Formula

In this section we present some results obtained by our "*fuzzy-based*" approach. The used link score computation formula is based on the "*weighted links approach*". In our experiments the λ parameter (equation V.7) has been set to 0.2, i.e., navigational link is five times important compared to a hierarchical link.

Table VI.19 presents iP[0.01] values obtained by our fuzzy approach compared to *Baseline*, *Topical PageRank*(M'hamed Mataoui et al., 2010) and *Demspter–Shafer*(M'hamed Mataoui, 2014)approaches results over the 107 content and structure queries of INEX 2007. We note from Table VI.19 that the *fuzzy approach* improves accuracy overall topics. Improvement is about 7.94% compared to *baseline*, 4.02% compared to *Topical Pagerank* and about 1.77% compared to *DS* approach for the case of Dalian information retrieval system. Comparison between fuzzy obtained results and baseline results for the three XML IR systems clearly shows that the fuzzy approach is the best with an improvement rate over 5%.

Figure VI.10 presents the obtained iP[0.01] values for some INEX 2007 topics (414, 415, 416, 419, 421, 424, 425, 426, 433, 434, 473 and 521) by our fuzzy approach compared to *baseline*, *Topical PageRank* and *Demspter–Shafer* approaches. We noticed that Fuzzy approach has improved most of the topics. For the topics in which improvement is equal to 0 (obtained iP[0.01] value equal to that of baseline), our approach obtained the highest value (iP[0.01] = 1) which confirms the importance of the value of the content evidence supported by the link evidence. For other topics of our experiments, e.g. 417 and 436, there was no improvement due to the lowest initial accuracy.

	Dalian IRS Retrieval results	Waterloo IRS retrieval results	MaxPlanck IRS retrieval results
Baseline	0.5271	0.5108	0.5066
Topical Pagerank	0.5470	0.5218	0.5072
DS without discounting rate	0.5682	0.5502	0.5310
DS with discounting rate	0.5591	0.5484	0.5281
Fuzzy Combination	0.5690 (7.94%*)	0.5510 (7.87%*)	0.5342 (5.44%*)
* % of Improvement compared to baseline results			

Table VI.19 : Obtained iP[0.01] values by our Fuzzy based approach comparedto baseline, Topical_Pagerankand DS based approach

The use of the "element-element" link type instead of "document-document" makes more sense to the improvement obtained by fuzzy-based approach compared to that of "*Topical-Pagerank*" approach.

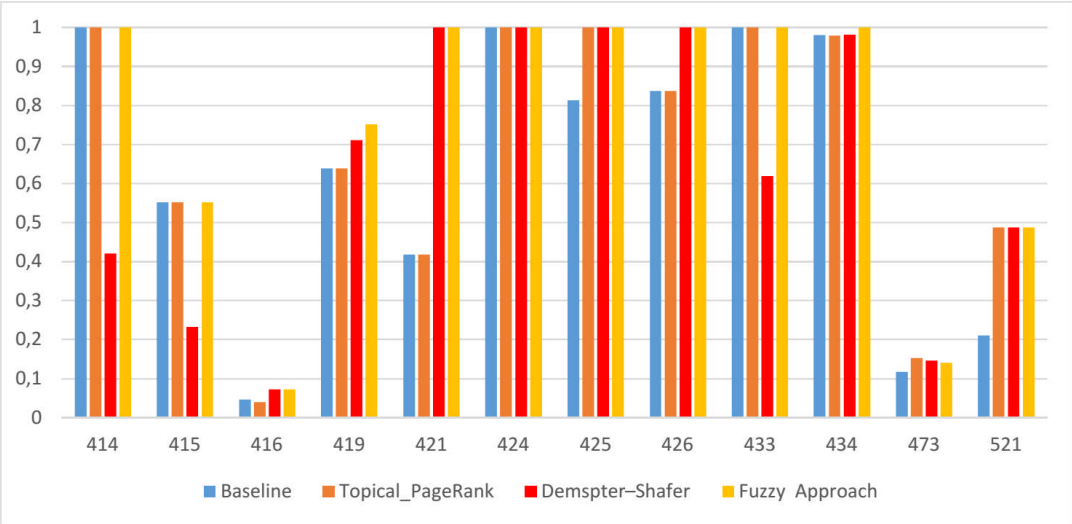


Figure VI.10: Obtained iP[0.01] values by our Fuzzy based approach compared to other approaches.

VI.5. Conclusion

We presented in this chapter the experimental setup, evaluation protocol and obtained results of our different proposals.

First, we have described the experimental setup including the details of test collections, tools and evaluation measures. Second, evaluation protocol which defined the way the experimental results have been obtained. Finally, obtained results have been described and discussed. During our experiments many features have been studied: the impact of internal links and external links; the impact of link score; the impact of the quality of initial retrieval results, etc.

Overall results show that:

- links represent a very important source of evidence in the information retrieval process;
- the proposed approaches have improved the retrieval accuracy for the different tested configurations;
- the "*topic-sensitive*" approaches are the best compared to global context approaches;
- approaches exploiting the links of type "element-element" obtained good results;
- The combination formulas taking into account the uncertain aspects of computed scores of XML elements allow achieving good performance.

Conclusion

Summary of the Thesis Work

The work presented throughout this thesis is in the context of information retrieval, particularly information retrieval in semi-structured XML documents. Currently, information retrieval has evolved from access to a document (or set of documents) to the access to information that meets particular interest of the user.

The effective exploitation of structured and semi-structured documents available requires to take into account the structural dimension of these documents. This structural dimension has led to the appearance of a new challenges in the information retrieval field. Indeed, the hierarchical structure of XML documents has been exploited as new evidence to retrieve XML elements at various levels of granularity. The XML structure has been used to provide a focused access to documents, by returning document component (e.g. sections, paragraphs, etc.), instead of whole document in response a user query.

The aim of our work is to propose approaches that exploit links as a source of evidence in the context of XML information retrieval.

The first issue we addressed concerns the possibility of considering the links as a relevance indicator. We have conducted a statistical study on the retrieval results returned by DALIAN information retrieval system. The performed study showed clearly that the links represent a sign of relevance of the returned XML elements in the context of XML IR using the INEX Wikipedia test collection.

We also adapted some Web link analysis algorithms (in our case: PageRank, HITS and SALSA) to XML Information Retrieval, particularly in the case of the INEX collection. We have implemented three Web link analysis algorithms, i.e. PageRank, HITS and SALSA, in the global context (Pagerank) and local context (Topical Pagerank, HITS and SALSA). We have proposed a new adaptation of the Pagerank algorithm, called "Topical Pagerank", by using it in the local context (topic-sensitive). We showed that the adaptation slightly outperform the baselines where the XML elements belonging to the top-20 documents are re-ranked. This "Topical Pagerank" obtained an improvement rate equal to 3.78%. Also, the SALSA implemented algorithm obtained an improvement rate of 7.92%. These implementations allows us to carry out a comparative study showing that query-dependent approaches (topic-sensitive) obtained the best performance compared to the query-independent (Pagerank) approaches.

We have proposed, implemented and evaluated a query dependent XML IR approach exploiting both intra document structure (hierarchical links) and external element-to-document (navigational) links to assign a link score to each retrieved element. We called this approach "Distance Based" approach. This approach is characterized by many features:

- First, it exploits element-to-document links to build element-to-element links for a given topic;
- Second, it exploits the two types of links: hierarchical and navigational links;
- Finally, it introduces the notion of "link distance" in the link score computation.

Contrary to the previous work, our approach exploit "element-element" link type, composed either by hierarchical and navigational links. Since most of XML collections contains "element-document" link type, we have proposed a solution that allows to propagate "element-document" link to the elements of the target document. Link score is computed only for retrieved elements. So, each retrieved element will propagate its relevance score towards its links to the other elements. The amount of propagated relevance score depends on the type of links (hierarchical and navigational) and the distance separating the source node and its target nodes. The "distance based" approach have been evaluated according to a linear combination formula with data extracted from INEX 2007 and 2009 retrieval results. Many features has been detailed, for instance, the impact of internal links, the impact of link score and the impact of the initial retrieval list quality. For the first feature, we have found that the external links are more important, in our context, than the internal links and that the use of the two link types is more effective compared to the use of only external links.

Concerning the second studied feature, i.e. , the impact of link score, Our main aim was to experiment the best value of the parameter that determines the degree of contribution of each score (initial or link scores) in the final score. The most important finding is that "distance based" approach outperforms the other approaches (previously cited) in all configurations. The paired-sample t-test show that our distance based approach is statistically significant compared to the baseline and other algorithms.

We have proposed also an alternative of our distance based approach. This Alternative approach, that we call "weighted links based" approach, is a "topic-sensitive" approach that combines both initial content relevance score and link evidence score to compute a new relevance score for each retrieved XML element. We have assumed that each retrieved XML element has a given relevance score that can be propagated through links. The amount of relevance score propagated between two XML elements, E_1 and E_2, is interpreted as the probability to explore this path by a user. The propagated amount of relevance is inversely proportional to the "path weight". Therefore, the more the path weight between two XML nodes is great, the more the probability to explore this path by a user is less.

The proposed approach used two combination formulas: The first combination formula is based on using the Dempster-Shafer theory of evidence (with and without discounting rate). It combines content relevance evidence for each retrieved XML element with its computed link evidence (according to "weighted links based" approach). The use of the Dempster–Shafer theory is motivated by the need to improve retrieval accuracy by incorporating the uncertain nature of both content and link relevance. The second combination formula we propose is based on fuzzy logic concepts.

To evaluate the "weighted links based" approach we carried out a set of experiments based on INEX 2007 test collection. We have noted that the two variants of the Dempster-Shafer combination formula (with and without discounting rate) improve the retrieval accuracy, and the variant without the discounting rate outperforms the one using the discounting rate. We have noted also from the obtained results that the fuzzy combination approach improves accuracy overall topics. Improvement is about 7.94% (DALIAN IR system) compared to baseline, 4.02% compared to Topical Pagerank and about 1.77% compared to DS approach for the case of Dalian information retrieval system.

Finally, we can mention that the use of link evidence allows through the proposal of several approaches (our link score computation and combination methods) to improve the retrieval accuracy by varying different parameters, with several types of links, by using two test collections (INEX 2007 and 2009) and in both contexts (global and local).

Perspectives

Further to this work, we can consider the following perspectives:

The first perspective is to exploit the proposed approaches on other test collections containing "element-element" link type as well as other retrieval tasks (Multimedia, book search, etc.) to study their behaviour while applied on other datasets.

The second perspective is the way link evidence is incorporated into the retrieval model. In our approaches we have only used computed link scores in combination with retrieval model scores (initial score) to re-rank. But link evidence and content evidence could be combined in different ways. In our case, link evidence is dependent on the ranking function used to obtain the top retrieved elements. An alternative is to propose another way in which local link evidence could be used directly in the ranking function of the retrieval model.

The third perspective concerns the combination formulas. As part of this thesiswe have experimented three combination formulas, i.e., linear, Dempster–Shaferand fuzzy based. Other forms of combination formulas can achieve better performance, i.e. by experimenting other discounting rate formulas for the Dempster–Shafer formula. For the case of fuzzy based combination formula more fuzzy sets might be defined if greater granularity is desired.

Finally,the perspective concerningthe choice of the XML elements of the topical sub graph. As HITS, which allows to take into account elements connected (but not in the first retrieved set) to the top-ranked elements, other methods may be proposed to take into account this aspect.

Bibliography

Adafre, S. F., & de Rijke, M. (2005). *Discovering missing links in Wikipedia*. Paper presented at the Proceedings of the 3rd international workshop on Link discovery.

Adamson, G. W., & Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information storage and retrieval, 10*(7), 253-260.

Agosti, M., Crestani, F., & Melucci, M. (1997). On the use of information retrieval techniques for the automatic construction of hypertext. *Information Processing & Management, 33*(2), 133-144.

Alashqur, A., Su, S. Y., & Lam, H. (1989). *OQL: a query language for manipulating object-oriented databases*. Paper presented at the Proceedings of the 15th international conference on Very large data bases.

Allan, J. (1997). Building hypertext using information retrieval. *Information Processing & Management, 33*(2), 145-159.

Anderson, J. D., & Pérez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management, 37*(2), 231-254.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463): ACM press New York.

Bharat, K., Broder, A., Dean, J., & Henzinger, M. R. (2000). A comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society for Information Science, 51*(12), 1114-1122.

Bharat, K., Chang, B.-W., Henzinger, M., & Ruhl, M. (2001). *Who links to whom: Mining linkage between web sites*. Paper presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on.

Bharat, K., & Henzinger, M. R. (1998). *Improved algorithms for topic distillation in a hyperlinked environment*. Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.

Bharat, K., & Mihaila, G. A. (2001). *When experts agree: Using non-affiliated experts to rank popular topics*. Paper presented at the Proceedings of the 10th international conference on World Wide Web.

Bharat, K. A., & Henzinger, M. R. (2000). Method for ranking documents in a hyperlinked environment using connectivity and selective content analysis: Google Patents.

Boughanem, M., Kraaij, W., & Nie, J.-Y. (2004). Modèles de langue pour la recherche d'information. *Les systemes de recherche d'informations*, 163-182.

Bratsas, C., Koutkias, V., Kaimakamis, E., Bamidis, P., & Maglaveras, N. (2007). *Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions.* Paper presented at the Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems, 30*(1), 107-117.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer networks and ISDN systems, 30*(1), 65-74.

Chamberlin, D., Robie, J., & Florescu, D. (2001). Quilt: An XML query language for heterogeneous data sources *The World Wide Web and Databases* (pp. 1-25): Springer.

Chamberlin, D. D., & Boyce, R. F. (1974). *SEQUEL: A structured English query language.* Paper presented at the Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control.

Chien, S., Dwork, C., Kumar, R., & Sivakumar, D. (2001). Towards exploiting link evolution.

Clark, J., & DeRose, S. (1999). XML path language (XPath).

Cleverdon, C. (1967). *The Cranfield tests on index language devices.* Paper presented at the Aslib proceedings.

Consortium, W. W. W. (1998). Extensible markup language (xml) 1.0. *W3C XML, February.*

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management, 36*(4), 533-550.

Crouch, C. (2005). *Relevance feedback at the INEX 2004 workshop.* Paper presented at the ACM SIGIR Forum.

DBLP. XML - dblp computer science bibliography – universitt trier. 2013, from http://dblp.uni-trier.de/xml/

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, 325-339.

Denoyer, L., & Gallinari, P. (2007). The wikipedia xml corpus *Comparative Evaluation of XML Information Retrieval Systems* (pp. 12-19): Springer.

Deutsch, A., Fernandez, M., Florescu, D., Levy, A., & Suciu, D. (1998). Xml-ql: A query language for xml: Citeseer.

Ding, C., He, X., Husbands, P., Zha, H., & Simon, H. D. (2002). *PageRank, HITS and a unified framework for link analysis.* Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Fachry, K. N., Kamps, J., Koolen, M., & Zhang, J. (2008). Using and detecting links in wikipedia *Focused access to XML documents* (pp. 388-403): Springer.

Farahat, A., LoFaro, T., Miller, J. C., Rae, G., & Ward, L. A. (2006). Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing, 27*(4), 1181-1201.

Florescu, D., & Kossmann, D. (1999). Storing and Querying XML Data using an RDMBS. *IEEE Data Eng. Bull., 22*(3), 27-34.

Frakes, W. B. (1992). Stemming Algorithms.

Fuhr, N., Gövert, N., & Großjohann, K. (2002). *HyREX: Hyper-media retrieval engine for XML*. Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Fuhr, N., & Großjohann, K. (2001). *XIRQL: A query language for information retrieval in XML documents*. Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

Fuhr, N., Kamps, J., Lalmas, M., Malik, S., & Trotman, A. (2008). Overview of the INEX 2007 ad hoc track *Focused Access to XML Documents* (pp. 1-23): Springer.

Fuller, M., Mackie, E., Sacks-Davis, R., & Wilkinson, R. (1993). *Structured answers for a large structured document collection*. Paper presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Geva, S. (2008). Gpx: Ad-hoc queries and automated link discovery in the wikipedia *Focused Access to XML Documents* (pp. 404-416): Springer.

Geva, S., Kamps, J., Lethonen, M., Schenkel, R., Thom, J. A., & Trotman, A. (2010). Overview of the INEX 2009 ad hoc track *Focused Retrieval and Evaluation* (pp. 4-25): Springer.

Geva, S., Kamps, J., & Trotman, A. (2009). INEX 2009 workshop pre-proceedings.

Geva, S., Trotman, A., & Tang, L.-X. (2009). Link Discovery in the Wikipedia. *Pre-Proceedings of INEX 2009*.

Gevrey, J., & Rüger, S. M. (2001). *Link-based Approaches for Text Retrieval*. Paper presented at the TREC.

Gövert, N., Abolhassani, M., Fuhr, N., & Großjohann, K. (2002). *Content-oriented XML retrieval with HyRex*. Paper presented at the INEX Workshop.

Gövert, N., & Kazai, G. (2002). *Overview of the Initiative for the Evaluation of XML retrieval (INEX) 2002*. Paper presented at the INEX Workshop.

Granitzer, M., Seifert, C., & Zechner, M. (2009). Context based wikipedia linking *Advances in Focused Retrieval* (pp. 354-365): Springer.

Grust, T. (2002). *Accelerating XPath location steps*. Paper presented at the Proceedings of the 2002 ACM SIGMOD international conference on Management of data.

Guo, L., Shao, F., Botev, C., & Shanmugasundaram, J. (2003). *XRANK: ranked keyword search over XML documents*. Paper presented at the Proceedings of the 2003 ACM SIGMOD international conference on Management of data.

Harman, D. (1993). *Overview of the first TREC conference*. Paper presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Harman, D. K. (1993). The first text retrieval conference (TREC-1) Rockville, MD, USA, 4–6 November, 1992. *Information Processing & Management, 29*(4), 411-414.

Haveliwala, T. (1999). Efficient computation of PageRank.

He, J., & de Rijke, M. (2010). *A ranking approach to target detection for automatic link generation*. Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

Henzinger, M. (2005). *Hyperlink analysis on the world wide web*. Paper presented at the Proceedings of the sixteenth ACM conference on Hypertext and hypermedia.

Henzinger, M. R. (2000). Link analysis in web information retrieval. *IEEE Data Eng. Bull., 23*(3), 3-8.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999). Measuring index quality using random walks on the Web. *Computer Networks, 31*(11), 1291-1303.

Hiemstra, D., & Mihajlovic, V. (2005). The simplest evaluation measures for XML information retrieval that could possibly work.

Hoffart, J., Bär, D., Zesch, T., & Gurevych, I. (2009). *Discovering Links Using Semantic Relatedness*. Paper presented at the Preproceedings of the INEX Workshop.

Huang, D. W. C., Xu, Y., Trotman, A., & Geva, S. (2008). Overview of INEX 2007 link the wiki track *Focused Access to XML Documents* (pp. 373-387): Springer.

INEX. Initiative for the Evaluation of XML Retrieval. 2013, from http://www.inex.otago.ac.nz/

Itakura, K. Y., & Clarke, C. L. (2008). University of waterloo at INEX2007: Adhoc and link-the-wiki tracks *Focused Access to XML Documents* (pp. 417-425): Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422-446.

Jia, X.-F., Alexander, D., Wood, V., & Trotman, A. (2011). University of Otago at INEX 2010 *Comparative Evaluation of Focused Retrieval* (pp. 250-268): Springer.

Kamps, J., Geva, S., Trotman, A., Woodley, A., & Koolen, M. (2009). Overview of the INEX 2008 ad hoc track *Advances in Focused Retrieval* (pp. 1-28): Springer.

Kamps, J., & Koolen, M. (2008). The importance of link evidence in Wikipedia *Advances in Information Retrieval* (pp. 270-282): Springer.

Kamps, J., & Koolen, M. (2009). *Is wikipedia link structure different?* Paper presented at the Proceedings of the Second ACM International Conference on Web Search and Data Mining.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., & Robertson, S. (2008). INEX 2007 evaluation measures *Focused access to XML documents* (pp. 24-33): Springer.

Kc, M., Chau, R., Hagenbuchner, M., Tsoi, A., & Lee, V. (2009). Link Prediction for Interlinked Documents by using Probability Measure Self Organizing Maps for Structured Domains. *Shlomo Geva, Jaap Kamps, Andrew Trotman*, 302.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation, 14*(1), 10-25.

Khokale, R. S., & Atique, M. (2013). Web Based Information Retrieval using Fuzzy Logic. *Int. J. of Soft Computing and Software Engineering, 3*(3), 62-68.

Kim, S., & Han, S. (2009). *The method of inferring trust in web-based social network using fuzzy logic.* Paper presented at the international workshop on machine intelligence research.

Kimelfeld, B., Kovacs, E., Sagiv, Y., & Yahav, D. (2007). Using language models and the HITS algorithm for XML retrieval *Comparative Evaluation of XML Information Retrieval Systems* (pp. 253-260): Springer.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604-632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods *Computing and combinatorics* (pp. 1-17): Springer.

Knoth, P., Novotny, J., & Zdrahal, Z. (2010). *Automatic generation of inter-passage links based on semantic similarity.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics.

Kohonen, T. (2001). *Self-organizing maps* (Vol. 30): Springer Science & Business Media.

Koolen, M. (2011). *The Meaning of Structure: the Value of Link Evidence for Information Retrieval.* Universiteit van Amsterdam, IR Publications, Amsterdam.

Koolen, M., Kazai, G., & Craswell, N. (2009). *Wikipedia pages as entry points for book search.* Paper presented at the Proceedings of the Second ACM International Conference on Web Search and Data Mining.

Kuropka, D. (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente-Information-Filtering und-Retrieval mit relationalen Datenbanken. *Advances in Information Systems and Management Science, 10*.

Lalmas, M. (2009). XML Retrieval (Synthesis Lectures on Information Concepts, Retrieval, and Services). *Morgan and Claypool, San Rafael, CA*.

Lalmas, M., Kazai, G., Kamps, J., Pehcevski, J., Piwowarski, B., & Robertson, S. (2007). INEX 2006 evaluation measures *Comparative Evaluation of XML Information Retrieval Systems* (pp. 20-34): Springer.

Lalmas, M., & Piwowarski, B. INEX 2005 Relevance Assessment Guide, 2005.

Lalmas, M., & Piwowarski, B. (2007). *INEX 2007 relevance assessment guide.* Paper presented at the Pre-Proceedings of inex.

Lalmas, M., & Ruthven, I. (1998). Representing and retrieving structured documents using the dempster-shafer theory of evidence: Modelling and evaluation. *Journal of documentation, 54*(5), 529-565.

Langville, A. N., & Meyer, C. D. (2005). A survey of eigenvector methods for web information retrieval. *SIAM review, 47*(1), 135-161.

Lee, C.-S., Wang, M.-H., & Hagras, H. (2010). A type-2 fuzzy ontology and its application to personal diabetic-diet recommendation. *Fuzzy Systems, IEEE Transactions on, 18*(2), 374-395.

Lee, Y. K., Yoo, S.-J., Yoon, K., & Berra, P. B. (1996). *Index structures for structured documents.* Paper presented at the Proceedings of the first ACM international conference on Digital libraries.

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks, 33*(1), 387-401.

Lempel, R., & Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS), 19*(2), 131-160.

M'hamed Mataoui, M. M., Faouzi Sebbak and Farid Benhammadi. (2014). *Evidential-Link-based Approach for Re-ranking XML Retrieval Results.* Paper presented at the 3rd International Conference on Data Management Technologies and Applications (DATA 2014), Vienna, Austria.

M'hamed Mataoui, F. S., Farid Benhammadi, Kadda Beghdad Bey. (2015). *A Fuzzy Link-Based Approach for XML Information Retrieval.* Paper presented at the The 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015), Istanbul, Turkey. http://fuzzieee2015.org/

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Marchiori, M. (1997). The quest for correct information on the web: Hyper search engines. *Computer networks and ISDN systems, 29*(8), 1225-1235.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM), 7*(3), 216-244.

Mass, Y., & Mandelbrod, M. (2003). *Retrieving the most relevant XML components.* Paper presented at the INEX 2003 Workshop Proceedings.

Mataoui, M. h., & Mezghiche, M. (2009). *Prise en compte des liens pour améliorer la recherche d'information structurée.*Paper presented at the CORIA.

Mataoui, M. h., & Mezghiche, M. (2015). A distance based approach for link analysis in xml information retrieval. *Computer systems science and engineering, 30*(3), 173-183.

Mataoui, M. h., Mezghiche, M., & Boughanem, M. (2010). Exploiting link evidence to improve XML information retrieval. *Conférence Internationale sur l'Extraction et la Gestion des Connaissances Maghreb (EGC-M 2010)*. doi: 10.13140/2.1.2014.6563

Mihajlović, V., Ramirez, G., De Vries, A. P., Hiemstra, D., & Blok, H. E. (2005). TIJAH at INEX 2004 modeling phrases and relevance feedback *Advances in XML Information Retrieval* (pp. 276-291): Springer.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with computers, 10*(3), 303-320.

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American documentation, 2*(1), 20-32.

Mulhem, P., & Verbyst, D. (2009). *Utilisation des liens entre documents structurés pour la recherche d'information.*Paper presented at the CORIA.

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). *Stable algorithms for link analysis.* Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

NIST. (2013). 2013 TREC Tracks. Retrieved 12/09/2013, 2013, from http://trec.nist.gov/tracks.html

Pehcevski, J., Vercoustre, A.-M., & Thom, J. A. (2008). Exploiting locality of Wikipedia links in entity ranking *Advances in Information Retrieval* (pp. 258-269): Springer.

Peters, C. (2001). *Cross-language information retrieval and evaluation*: Springer.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems, 14*(3), 130-137.

Rafiei, D., & Mendelzon, A. O. (2000). What is this page known for? Computing Web page reputations. *Computer Networks, 33*(1), 823-835.

Ramírez, G., Westerveld, T., & de Vries, A. P. (2005). *Structural features in content oriented XML retrieval.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Rayward, W. B. (1994). Visions of Xanadu: Paul Otlet (1868–1944) and hypertext. *Journal of the American Society for Information Science, 45*(4), 235-250. doi: 10.1002/(SICI)1097-4571(199405)45:4<235::AID-ASI2>3.0.CO;2-Y

Rijsbergen, C. J. v. (1979). *Information Retrieval*: Butterworth.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of documentation, 33*(4), 294-304.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129-146.

Robertson, S. E., & Walker, S. (1994). *Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval.* Paper presented at the Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP*, 109-109.

Robie, J., Derksen, E., Fankhauser, P., Howland, E., Huck, G., Macherius, I., . . . Schöning, H. (1999). XQL (XML query language). *Online only*.

Saint Laurent, S., & Petitjean, A. (2000). *Introduction au XML*: Osman Eyrolles multimédia.

Salton, G. (1971). The SMART retrieval system—experiments in automatic document processing.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM, 26*(11), 1022-1036.

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*(6), 321-343.

Sauvagnat, K. (2005). *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés.* (Phd Thesis), Université Paul Sabatier, Toulouse, France.

Sauvagnat, K., & Boughanem, M. (2006). *Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée.*Paper presented at the CORIA.

Sauvagnat, K., Hlaoua, L., & Boughanem, M. (2006). XFIRM at INEX 2005: ad-hoc and relevance feedback tracks *Advances in XML Information Retrieval and Evaluation* (pp. 88-103): Springer.

Schmidt, A., Waas, F., Kersten, M., Carey, M. J., Manolescu, I., & Busse, R. (2002). *XMark: A benchmark for XML data management.* Paper presented at the Proceedings of the 28th international conference on Very Large Data Bases.

Schocken, S., & Hummel, R. A. (1993). On the use of the Dempster Shafer model in information indexing and retrieval applications. *International Journal of Man-Machine Studies, 39*(5), 843-879.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 1): Princeton university press Princeton.

Shaw Jr, W. M., Burgin, R., & Howell, P. (1997). Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing & Management, 33*(1), 1-14.

Sigurbjörnsson, B., Trotman, A., Geva, S., Lalmas, M., Larsen, B., & Malik, S. (2005). *INEX 2005 guidelines for topic development.* Paper presented at the INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing & Management, 32*(5), 619-633.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Standardization, I. O. f. (1986). *Information Processing: Text and Office Systems: Standard Generalized Markup Language (SGML)* (Vol. 8879): ISO Geneva.

Theobald, A., & Weikum, G. (2002). The index-based XXL search engine for querying XML data with relevance ranking *Advances in Database Technology—EDBT 2002* (pp. 477-495): Springer.

Trotman, A., & Sigurbjörnsson, B. (2005). Narrowed extended xpath i (nexi) *Advances in XML Information Retrieval* (pp. 16-40): Springer.

Turtle, H., & Croft, W. B. (1989). *Inference networks for document retrieval.* Paper presented at the Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval.

Van Zwol, R., Kazai, G., & Lalmas, M. (2006). INEX 2005 multimedia track *Advances in XML Information Retrieval and Evaluation* (pp. 497-510): Springer.

Verbyst, D., & Mulhem, P. (2009). Using Collectionlinks and Documents as Context for INEX 2008 *Advances in Focused Retrieval* (pp. 87-96): Springer.

Wikipedia. Wikipedia: The free encyclopedia. Retrieved October 2013, 2013, from http://en.wikipedia.org/

Wong, S. M., Ziarko, W., & Wong, P. C. (1985). *Generalized vector spaces model in information retrieval.* Paper presented at the Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval.

Wood, L., Le Hors, A., Apparao, V., Byrne, S., Champion, M., Isaacs, S., . . . Sutor, R. (1998). Document object model (DOM) level 1 specification. *W3C Recommendation, October*.

Yang, K. (2005). Information retrieval on the web. *ARIST, 39*(1), 33-80.

Zhang, J., & Kamps, J. (2008). *Link detection in XML documents: What about repeated links.* Paper presented at the SIGIR 2008 Workshop on Focused Retrieval.

Appendix

Appendix A: Link detection - state of the art (continued)-

In (Fachry et al., 2008), authors describe their participation in the Link the Wiki track of INEX 2007. Authors investigated the relative effectiveness of link detection based only on the Wikipedia article's name, and on the matching arbitrary text segments of different pages. Their approach search the Wikipedia pages for some text segments shared between two XML nodes to detect that they are implicitly connected. The proposed approach is mostly based on the assumption that the shared text segments is only one specific string (Agosti et al., 1997). One text segment as a single line is defined and a string that the two XML nodes share is considered as a relevant substring.

Only relevant substrings of at least 3 characters length are considered in the approach to prevent detecting too many false positives. It assumes that XML documents that link to each other are someway semantically related. Khairun et al (Fachry et al., 2008) adopt a breadth m–depth n technique for automatic text structuring for identifying candidate anchors and text node. The similarity on the document level and text segment level is used as evidence an then as a precision filter. The proposed approach consists of two steps: First, it detects links on the XML document level. It retrieve the top N similar XML documents in the collection by using the whole orphan XML document (without structure) as a query. Second, it detect links on the XML element level. It search on the local level with text segments. Normalized lines are matched with string processing and keep track of the absolute path for each text node and calculate the offset (starting and end position) of the identified anchor text.

The defined Best Entry Point (BEP) for both incoming and outgoing links was the start of an XML document ("*/article[1]/name[1]*" element). Assuming that links are not reciprocal, different approaches for detecting outgoing and incoming links are proposed, though a threshold of 250 for both type of links is set. As requested in the INEX LTW task specification duplicated links are not allowed.

To detect "Outgoing Links" which represent links from an anchor text in the topic file to the BEP of existing semantically related XML documents in the collection, in this case the text-node of the "*/article[1]/name[1]*" element. There is an outgoing link for topic t, when $S_{1\cdots n} = T_{q\cdots r}$, where S is the title of a target XML document (foster article), and T is a line in an orphan article. For detecting "Incoming Links" which represent links consisting of a specified anchor connecting text nodes in the target articles to the "/article[1]/name[1]" element of one of orphan XML documents. There is an incoming link for topic t, when $T_{1\cdots n}=S_{q\cdots r}$, where $T_{1\cdots n}$ is the title of t, and S is a line in a foster article.

From each topic, the title enclosed with the <name> tag is extracted with a regular expression. Only "article-to-article" links are considered in the scores. The threshold for the number of incoming and outgoing links is set to 250 for each topic.

Best performance are achieved by setting the threshold of the result list to the top 300 retrieved XML documents (MAPin =0.3713). Results shows that while the recall improves, which has slight positive effect on the performance for the outgoing links, the precision drops and thus the fallout also increases for the incoming links. In (Fachry et al., 2008), Vector Space Model is

experimented with the entire orphaned XML documents as query. Results show that exact substring matching improves the performance as compared to generating plain "article-to-article" links.

	Outgoing			Incoming		
Run	**MAP**	**R-Prec**	**P@5**	**MAP**	**R-Prec**	**P@5**
Article100	0.1518	0.2277	0.5711	0.2646	0.3062	0.7311
Name100	0.1533 +1.0%	0.1781	0.7489	0.2906 +9.8%	0.3134	0.8000
Article200	0.1629	0.2389	0.5711	0.3075	0.3529	0.7311
Name200	0.1739 +6.8%	0.2073	0.7356	0.3471 +12.9%	0.3835	0.8044
Article250	0.1658	0.2406	0.5711	0.3193	0.3628	0.7311
Name250	0.1783 +4.9%	0.2147	0.7267	0.3618 +13.3%	0.3998	0.8044
Article300	0.1678	0.2407	0.5711	0.3274	0.3691	0.7311
Name300	0.1825 +8.8%	0.2233	0.7178	**0.3713** +13.4%	0.4101	0.8044
Name400	**0.1836**	0.2405	0.6844	0.3117	0.3757	0.6067

Table A.1 : Results for the Link The Wiki track (Fachry et al., 2008)

The proposed approach focuses on exact string matching, but have not explored best matching techniques such as semantic clustering of words, which could further improve the performance.

In (Zhang & Kamps, 2008), authors provide an extensive analysis of two aspects of links: density and repetition. They study the impact of some parameters like the document's length, the distance between anchor text occurrences, and the frequency of the anchor text within an article. They start the study by considering that excessive links make a Wikipedia article difficult to read and taking the hypothesis that good links in Wikipedia are relevant to the context. Therefore, adding more links to an XML document improve its context. In the context of (Zhang & Kamps, 2008) research, three questions have been addressed. For the first question: Does link repetition occur, and how often? Authors showed that the same links do re-occur in the same documents. A considerable subset of the number of links are actual repeated links. For the second question: How can we predict when to link in a XML document? Authors present Vector Space Model to cluster related documents and then find relevant anchor terms. In addition, they extract "<name>" XML nodes relevant substrings of XML documents to consider them as source of links. Authors conducted a study with 2 variables, i.e. distance between 2 of the same anchor terms, and the total number of possible link candidates in a file, to predict when a repeated link candidate should be actually made a link. For the third question: Will link detection in XML documents improve by taking into account repetitions of links? Authors compared their runs with 2 baselines. Preliminary experiments showed that taking into account repeated links in the `ground truth' can achieve better link detection performance compared to the baselines of `linking once' and `link always'. Main finding of (Zhang & Kamps, 2008) is that, although the overall impact of link repetition is modest, performance can increase by taking an informed approach to link repetition.

Michael Granitzer et al. (Granitzer, Seifert, & Zechner, 2009) have outlined methods based on the context evidence, i.e. the surrounding text of a link candidate, for detecting links between Wikipedia XML documents automatically and evaluate the potential of different context types (different ranges of words surrounding the anchor) to calculate the relevance of a possible link candidate. The proposed approach focuses on a content based strategy by detecting documents titles as link candidates and selecting the most relevant ones as links. The presented work aims at evaluating the influence of the context on selecting relevant links and determining a

links best-entry-point. Link candidates matching are ranked using different context types. Authors (Granitzer et al., 2009) present a detailed study of parameters for estimating their influence on the context choice.

The implemented system defines a set of possible incoming links and a set of possible outgoing links for each orphan XML document. It determine the anchor context of each link where the nouns are extracted and fed into the retrieval backend as Boolean OR query. For acceleration, the system restrict the result set to XML documents pointed to by all links in the anchor context by adding all link target identifiers as AND query part. The proposed query is formulated as:

$$(ID = t_1 \ \ OR \ldots OR \ \ ID = t_n) \ AND \ (w_1 \ \ OR \ \ w_2 \ \ \ldots \ \ OR \ \ w_k) \tag{e.1}$$

Where w_i represents the nouns of the anchor context and t_k is unique identifier for the k^{th} link target. ID specifies the search on the metadata field containing the unique identifiers of a XML document of the LTW collection. The anchor context score is obtained by the following formula:

$$s_i = coord_{w,i} * norm(w) * \sum_{t \in w} \frac{\sqrt{tf_{t,i}} * idf_t^2}{norm(i)} \tag{e.2}$$

Where:

- $tf_{t,i}$ is the frequency of the term t in the XML document i.
- idf_t is the inverse document frequency.
- *norm(w)* is the norm of the query.
- *norm(i)* is the number of terms of the XML document i.
- $coord_{w,i}$ is the overlapping factor.

For the incoming links detection, authors (Granitzer et al., 2009) use the same principle by title matching by taking the best source candidates for determining the BEP or the file-to-file links. These BEP are determined again based on the link context. The hypothesis taking here is that the BEP in the link target has to be similar to the anchor context. As result, if the title of the source XML document is contained in the link target, those parts of the target document are preferred entry points. The best five entry points are taken as result.

Experiments are evaluated on the INEX 2007 LTW Wikipedia Corpus consisting of 659,413 XML (splited into two test sets). 2 runs for the file-to-file task have been submitted for comparing the best anchor context method with the context free anchor IDF approach. For anchor-to-bep a combination of different outgoing links and incoming links generation approaches are submitted by distinguishing between context based and context free approaches (Table A.2). Results show that the choice of the type of context is critical. The whole document seems to be best suited as anchor context, followed by automatically detected topics. Other XML document structures such as sections and paragraph are assessed as bad context choices which decrease detection accuracy. The proposed approach in (Granitzer et al., 2009) have achieved around 4% of increases precision by using the context evidence.

Parameters		file-to-file			anchor-to-bep
Title as OR Query	context	MAP_{intern}	$MAP_{official}$	$MAP_{reevaluated}$	$MAP_{official}$
false	document	0.6355	0.5300	0.625	0.2384
false	sentence	0.5938	NA	NA	0.1895
false	topic	NA	NA	NA	0.2619
false	no context	0.5938	0.5369	0.606	0.1968
true	document	0.5066	NA	NA	NA
true	sentence	0.4088	NA	NA	NA
true	no context	0.4088	NA	NA	NA

Table A.2 : Results for inlink generation file-to-file (Granitzer et al., 2009)

In (Kc, Chau, Hagenbuchner, Tsoi, & Lee, 2009), authors explain how Self-Organizing Maps (SOM) (Kohonen, 2001) can be used for link detection. They indicate that the self-organizing maps training algorithm encodes existing relations between the atomic elements in a graph. The proposed link detection approach is presented as scalable in fact that links can be extracted in linear time.

An XML structure (tree) can be considered as a special type of graphs. To encode this structure using a SOM, authors consider each XML node (subtree) in the graph as independent. Using hierarchical link type, data is processed from the leaf nodes to the XML document root node. To connect the SOM model of the different XML nodes, a method consisting of using the winning nodes and connect them to their parent nodes is proposed.

Training data has been processed by using the Probability Measure Graph SOM (PM-GraphSOM) which allows introducing cyclic dependencies (link information) by a SOM. A selection of features to be used as node labels for each XML document must be defined. Authors incorporate link information (136,304,216 existing links) to assist in the training process. Also, they used the Multi-Layer Perceptron (MLP) algorithm to assist in dimension reduction of the processed XML data (2,661,190 XML documents belonging to 362,251 unique categories).

An analysis of the outgoing and incoming was carried out. This analysis indicated that the number of incoming links varies much more than the number of outgoing links. This feature is important since it implies that the dataset is unbalanced regarding the incoming and outgoing links. According to (Kc et al., 2009), such unbalances in the training dataset can cause problems when using a machine learning approach. For this reason, authors (Kc et al., 2009) do not use link structure information during the training.

Three ranking algorithms were considered for the proposed learning for detecting links. The first is based on the energy flow of an XML document of the collection. This energy is computed by accumulating scores received by an XML document from different incoming links and then distributed through the different outgoing links. The second ranking algorithm is based on incoming and outgoing link frequencies. The third algorithm, based on the Euclidean distance between the test document and the training documents, ranks higher the XML documents of the training documents which are more similar.

Table A.3 presents recall and precision values obtained by the five proposed configurations (submissions). Each of the five submissions use combinations of training configuration and ranking algorithm.

Submission ID	01	02	03	04	05
Link inference method	Consider all	Content-based	Content-based	Codebook-based	Content-based
Ranking algorithm	Energy flow	Link frequency	Link frequency	Euclid. distance	Euclid. distance
Recall for out-links	0.00377	0.02942	0.03202	0.00070	0.00448
Precision@250 out-links	0.00136	0.01057	0.01817	0.00025	0.00161
Precision@100 out-links	0.00168	0.01795	0.02162	0.00002	0.00147
Precision@20 out-links	0.00040	0.03941	0.04489	0.00003	0.00268
Recall for in-links	0.00032	0.00050	0.00095	0.00010	0.00029
Precision@250 in-links	0.00008	0.00012	0.00025	0.00002	0.00007
Precision@100 in-links	0.00007	0.00008	0.00037	0.00003	0.00008
Precision@20 in-links	0.00007	0.00018	0.00107	0.00001	0.00008

Table A.03 : The performance of proposed links as predicted by our proposed algorithms (Kc et al., 2009)

In (Hoffart, Bär, Zesch, & Gurevych, 2009), authors present two unsupervised link detection approaches. The first approach uses collections containing existing links and the second approach assumes that the collection must be without existing links. In collections containing links, authors propose to use link anchor ranking measures based on existing links to detect new links. In collections without links, noun phrases are gathered to produce anchor candidates. In the target detection step, authors exploit the measure of semantic relatedness between texts. The detected links are of two levels: document-level links and anchor-level links. Document-level links consist of relating source documents to target ones. Anchor-level links consist of relating an anchor phrase in the source document to the target document by specifying the entry point (the root or any element of the document).

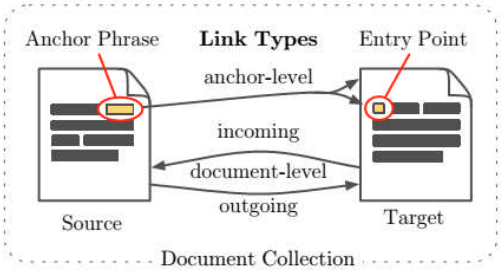


Figure A.1: Link types in documents collections (Hoffart et al., 2009)

Authors describe their approach to proceed with the two tasks of anchor-level link detection. For the first task, i.e. anchor discovery, they identify and rank the anchor candidates, which can be N-Grams (used by Geva (Geva, 2008)), Noun phrases or titles (document or section). Existing links are used in this step to improve the anchors ranking. For the second task, i.e. target discovery, the best matching target is retrieved by using anchor phrase as query. Link detection on document level is considered as generalization of the anchor-level detection approach.

To detect outgoing links, Hoffart et al. (Hoffart et al., 2009) identify potential anchors and their appropriate target, by using existing links. They combine the *keyphraseness* and as_{max} measure to rank anchor candidates and ESA semantic relatedness measure to disambiguate between potential targets. To detect incoming links, authors (Hoffart et al., 2009) execute a full-text search for the article title using the standard Lucene retrieval model, and take the top 250 results as sources of incoming links. Another approach consists on re-ranking the top 2000 retrieval results using ESA semantic relatedness measure between the orphan document and the potential source.

Geva et al. (Geva, Trotman, & Tang, 2009) describe link detection approaches presented in the LTW track of INEX campaign for the three tasks: *Link-the-Wiki*, *Link-Te-Ara*, and *Link-Te-Ara-to-the-Wiki*. They use two methods of outgoing links generation: the top ranking algorithms from previous LTW tracks (Itakura's *ICLM* algorithm and Geva's *GPNM* algorithm) and a hybrid method derived from them. Geva et al. (Geva, Trotman, et al., 2009) attempt to improve the performance by combining the scores computed according to both algorithms:

$$Score(L)=Score_s(L) + Score_k(L)$$

Where $Score_S(L)$ is the score from the GPNM algorithm and $Score_K(L)$ is a normalized ICLM for the link L.

For incoming links, Geva et al. use traditional information retrieval strategy on the Wikipedia XML collection by taking the top ranked results (250 pages) returned using the topic title as a query from a BM25 search engine. To find the Best-Entry-Points, Geva et al. (Geva, Trotman, et al., 2009) propose two approaches. The first approach consists of finding the first location of the anchor terms according to one of the two methods: exact matching or no exact matching. The second approach is similar to that used in image matching, i.e. images with more similar features describe similar objects. The proposed hybrid method for outgoing link detection doesn't work as well as the original ICLM algorithm

Xiang-Fei Jia et al. (Jia, Alexander, Wood, & Trotman, 2011) propose a new approach based on the simpler relevance summation method for producing best entry points as a solution that still an unsolved problem, i.e. link detection in a corpus that has no existing links. They argue the use of the Te Ara collection by the fact that it does not possess characteristics of Wikipedia collection that have become too easy for algorithms to score highly according to the metrics used by INEX. Te Ara collection documents are qualified as less "to-the-point" than Wikipedia documents (see section IV.2: Approaches studying of structure and nature of links in Wikipedia).Therefore, Xiang-Fei Jia et al. (Jia et al., 2011) judged significant to take into account the immediate context of a candidate anchor or entry-point, as well as the more general content of the two documents being linked. Authors has indexed three types of information: documents, sections and headings of sections which make it possible to vary the level of target context when searching for possible entry-points (targets). In the proposed approach every sequence of successive words without stopwords or punctuation separation marks are considered as an anchor candidate. The approach compute a relevance score for each target using BM25 retrieval formula for the different contexts' queries. Xiang-Fei Jia et al. (Jia et al., 2011) submitted 24 runs produced by varying the 4 parameters: full document anchor context, relevance summation method, relevance score contribution and target contexts.

In (Knoth, Novotny, & Zdrahal, 2010), authors consider the use of semantic similarity measures for link detection at document and passage levels. They studied relation between semantic similarity and the length of the documents properties. The study confirms that the length of documents is an important factor usually causing the quality of current link generation approaches to deteriorate. The proposed link detection strategy is formalized as a two-step process: firstly, identifying candidate pairs of documents that should be linked; then, identifying pairs of passages, for each candidate pairs of documents, for which topics are semantically related in both documents. The work presented in (Knoth et al., 2010) makes two principal contributions: providing a new interpretation of the use of semantic similarity measures for link detection; and developing a novel two-step method for detecting passage-to-passage links.

Authors indicate, as result, that high similarity score between two documents is not necessarily a good predictor for link detection process. They argue that by citing the case in

which users create links connecting documents that provide new information and not necessarily connect similar content documents.

The document-to-document link detection algorithm (Figure A.2) iterates over all pairs of document and return all document vector pairs with similarity higher than α and smaller than β parameters. To rank the detected links according to the confidence of the system, it is suggested to assign each pair of documents a score using the equatione.3.

Input: A set of document vectors D,
min. sim. α, max. sim. $\beta \in [0,1], C = \emptyset$
Output: A set C of candidate links
of form $\langle d_i, d_j, sim\rangle \in C$ where d_i and d_j are
documents and $sim \in [0,1]$ is their similarity
1.**for each** $\{\langle d_i, d_j\rangle | i,j \in \aleph_0 \wedge i < j < |D|\}$ **do**
2. $sim_{d_i,d_j} := similarity(d_i, d_j)$
3. **if** $sim_{d_i,d_j} > \alpha \wedge sim_{d_i,d_j} < \beta$ **then**
4. $C := C \cup \langle d_i, d_j, sim_{d_i,d_j}\rangle$

Figure 0A.2: Document-to-document link detection algorithm

Obtained results for this link detection algorithm clearly indicate that the document-to-document detection achieved high performance when parameters α, β are well selected.

$$rank_{d_i,d_j} = |sim_{d_i,d_j} - (\alpha + \frac{\beta - \alpha}{2})| \tag{0e.3}$$

For Passage-to-passage link type, links are detected using a two-step process by considering that the similarity between two passages is mostly higher than the similarity between documents to which these passages belong. Authors use this assumption to propose algorithm of Figure A.3 to set γ and δ parameters. Their algorithm indicates how "passage-to-passage" links are calculated for a single document pair identified previously by algorithm of Figure A.2**Erreur ! Source du renvoi introuvable.**. The two-step process allows detecting document pairs likely to contain strongly connected passages and to identify the related passages.

Input: Sets P_i, P_j of paragraph document
vectors for each pair in C
min. sim. γ, max. sim. $\delta \in [0,1]$ such that
$\alpha < \gamma \wedge \beta < \delta, , L = \emptyset$
Output: A set L of passage links
of form $\langle p_{k_i}, p_{l_j}, sim\rangle \in L$ where p_{k_i} and
p_{l_j} are paragraphs in documents d_i, d_j
and $sim \in [0,1]$ is their similarity
1.**for each** $\{\langle p_{k_i}, p_{l_j}\rangle | p_{k_i} \in P_i, p_{l_j} \in P_j\}$ **do**
2. $sim_{p_{k_i},p_{l_j}} := similarity(p_{k_i}, p_{l_j})$
3. **if** $sim_{p_{k_i},p_{l_j}} > \gamma \wedge sim_{p_{k_i},p_{l_j}} < \delta$ **then**
4. $L := L \cup \langle p_{k_i}, p_{l_j}, sim_{p_{k_i},p_{l_j}}\rangle$

FigureA.3: Generate passage links Algorithm

In evaluation section of (Knoth et al., 2010), authors have extracted pairs of documents by fixing $\alpha = 0.65$ and $\beta = 0.70$. The performed study revealed that 91% of the detected links were assessed as relevant and 9% as irrelevant.

In (He & de Rijke, 2010), authors focuses automatic link detection with Wikipedia collection and formulates the task of finding link targets as a ranking problem. They investigate the effectiveness of approaches based on learning to rank principal.

For this task of detecting link targets, authors propose to use N-gram techniques to find the related concepts in Wikipedia collection and explore many features that do not be dependent of a "non-ambiguous" context of a given N-gram. The used features can be categorised into three types:

- **N-gram-target features: which** describe how well an N-gram *ng* and a candidate target *ctar* are related. Three features are explored in this feature type:
  - (i) TitleMatch;
  - (ii) Link Evidence;
  - (iii) Retrieval scores
- **Target features: which contains f**our features:
  - (i) number of inlinks;
  - (ii) number of outlinks;
  - (iii) associated Wikipedia categories;
  - (iv) generality
- **Topic-Target features: which** describe the relatedness between a topic t and candidate target *ctar*. Two features are explored in this feature type:
  - (i) cosine similarity between t and *ctar*;
  - (ii) retrieval score using title of *ctar* as query and t as target document by using the BM25 retrieval model.

For experimental evaluation, authors (He & de Rijke, 2010) have used INEX 2008 Wikipedia collection to construct the training, validation and test sets. The training set contains 11,112 anchor texts, the validation set contains 9,365 anchors texts and 3,452 anchor texts in the test set. Three binary classifiers, namely *NaiveBayes*, *SVM* and *J48*, were used as baselines.

The performed experiments using the INEX 2008 Wikipedia collection show that learning to rank approaches (AdaRank of Table A.4) using these features significantly outperform binary classification approaches. Also, the proposed features can be well used for both binary classification and learning to rank settings. Results achieved by these features are as good as to other binary state-of-the-art approaches.

Method	MAP	p@1	p@5
Ranking SVM	0.9502	0.9235	0.1970
AdaRank	**0.9629**	**0.9395**	**0.1980**
SVM-class	$0.9476^{\triangledown\blacktriangledown}$	$0.9180^{\triangledown\blacktriangledown}$	$0.1970^{\blacktriangledown}$
J48	$0.9496^{\blacktriangledown}$	$0.9218^{\triangledown\blacktriangledown}$	$0.1968^{\blacktriangledown}$
NaiveBayes	$0.9200^{\triangledown\blacktriangledown}$	$0.8699^{\triangledown\blacktriangledown}$	$0.1962^{\triangledown\blacktriangledown}$

Table A.4 : Performance of learning to rank approaches compared to binary classification approaches (He & de Rijke, 2010)

Appendix B: Statistics

Table B.1. Percentage of link relevance (107 CO topics of INEX 2007, with top 20 relevant documents retrieved)

	IN			OUT		
TOPIC_ID	Relevant links	All links	% relevant	Relevant links	All links	% relevant
topic : 414	2	26	7,69	2	9	22,22
topic : 415	1	20	5,00	1	3	33,33
topic : 416	2	5	40,00	2	9	22,22
topic : 417	1	1	100,00	1	2	50,00
topic : 418	1	12	8,33	1	6	16,67
topic : 419	33	40	82,50	33	35	94,29
topic : 420	16	27	59,26	16	26	61,54
topic : 421	8	26	30,77	8	23	34,78
topic : 422	5	5	100,00	5	20	25,00
topic : 423	24	30	80,00	24	33	72,73
topic : 424	55	59	93,22	55	56	98,21
topic : 425	100	128	78,13	100	106	94,34
topic : 426	21	40	52,50	21	27	77,78
topic : 427	5	22	22,73	5	27	18,52
topic : 428	27	35	77,14	27	34	79,41
topic : 429	42	74	56,76	42	58	72,41
topic : 430	59	64	92,19	59	61	96,72
topic : 431	30	49	61,22	30	39	76,92
topic : 433	5	8	62,50	5	12	41,67
topic : 434	48	58	82,76	48	56	85,71
topic : 435	19	22	86,36	19	20	95,00
topic : 436	0	0	0,00	0	0	0,00
topic : 439	0	0	0,00	0	0	0,00
topic : 440	57	59	96,61	57	61	93,44
topic : 441	15	22	68,18	15	19	78,95
topic : 444	0	3	0,00	0	6	0,00
topic : 445	58	75	77,33	58	63	92,06
topic : 446	6	25	24,00	6	26	23,08
topic : 447	10	24	41,67	10	12	83,33
topic : 448	16	23	69,57	16	22	72,73
topic : 449	55	55	100,00	55	55	100,00
topic : 450	2	3	66,67	2	3	66,67
topic : 453	0	1	0,00	0	1	0,00
topic : 454	46	53	86,79	46	63	73,02
topic : 458	53	63	84,13	53	56	94,64

topic : 459	24	33	72,73	24	34	70,59
topic : 461	2	2	100,00	2	2	100,00
topic : 463	9	22	40,91	9	18	50,00
topic : 464	37	46	80,43	37	39	94,87
topic : 465	26	60	43,33	26	51	50,98
topic : 467	0	0	0,00	0	0	0,00
topic : 468	0	1	0,00	0	2	0,00
topic : 469	37	44	84,09	37	49	75,51
topic : 470	20	27	74,07	20	32	62,50
topic : 471	7	9	77,78	7	20	35,00
topic : 472	15	25	60,00	15	35	42,86
topic : 473	14	20	70,00	14	19	73,68
topic : 474	14	38	36,84	14	38	36,84
topic : 475	27	48	56,25	27	65	41,54
topic : 476	28	40	70,00	28	33	84,85
topic : 477	17	25	68,00	17	20	85,00
topic : 478	1	3	33,33	1	4	25,00
topic : 479	10	19	52,63	10	22	45,45
topic : 480	6	12	50,00	6	11	54,55
topic : 481	0	6	0,00	0	7	0,00
topic : 482	83	83	100,00	83	83	100,00
topic : 483	45	68	66,18	45	67	67,16
topic : 484	3	7	42,86	3	3	100,00
topic : 485	25	31	80,65	25	30	83,33
topic : 486	15	26	57,69	15	24	62,50
topic : 487	3	10	30,00	3	13	23,08
topic : 488	44	58	75,86	44	50	88,00
topic : 489	15	32	46,88	15	25	60,00
topic : 490	26	47	55,32	26	38	68,42
topic : 491	0	2	0,00	0	1	0,00
topic : 495	0	5	0,00	0	8	0,00
topic : 496	15	27	55,56	15	23	65,22
topic : 497	7	23	30,43	7	20	35,00
topic : 498	9	31	29,03	9	15	60,00
topic : 499	4	4	100,00	4	4	100,00
topic : 500	4	18	22,22	4	14	28,57
topic : 502	50	54	92,59	50	58	86,21
topic : 503	11	49	22,45	11	34	32,35
topic : 505	42	71	59,15	42	76	55,26
topic : 506	29	41	70,73	29	41	70,73
topic : 507	1	1	100,00	1	3	33,33
topic : 508	8	22	36,36	8	20	40,00
topic : 509	61	61	100,00	61	61	100,00
topic : 511	0	0	0,00	0	0	0,00
topic : 515	1	11	9,09	1	10	10,00

topic : 516	3	7	42,86	3	3	100,00
topic : 517	1	9	11,11	1	7	14,29
topic : 518	50	79	63,29	50	79	63,29
topic : 519	10	11	90,91	10	10	100,00
topic : 520	27	35	77,14	27	30	90,00
topic : 521	16	29	55,17	16	26	61,54
topic : 522	11	18	61,11	11	20	55,00
topic : 523	18	29	62,07	18	38	47,37
topic : 525	19	28	67,86	19	22	86,36
topic : 526	33	91	36,26	33	52	63,46
topic : 527	14	60	23,33	14	38	36,84
topic : 528	76	82	92,68	76	77	98,70
topic : 529	18	36	50,00	18	21	85,71
topic : 530	54	61	88,52	54	60	90,00
topic : 531	3	22	13,64	3	8	37,50
topic : 532	0	15	0,00	0	6	0,00
topic : 533	14	25	56,00	14	18	77,78
topic : 534	43	46	93,48	43	47	91,49
topic : 535	12	13	92,31	12	12	100,00
topic : 536	36	43	83,72	36	52	69,23
topic : 537	15	38	39,47	15	22	68,18
topic : 538	0	1	0,00	0	3	0,00
topic : 539	7	12	58,33	7	9	77,78
topic : 540	0	0	0,00	0	0	0,00
topic : 541	12	21	57,14	12	17	70,59
topic : 542	13	40	32,50	13	29	44,83
topic : 543	8	31	25,81	8	21	38,10
Total	2130	3301	64,53	2130	2998	71,05

Table B.2. Percentage of link relevance (107 CO topics of INEX 2007, with top 50 relevant documents retrieved)

	IN			OUT		
TOPIC_ID	Relevant links	All links	% relevant	Relevant links	All links	% relevant
topic : 414	5	65	7,69	5	31	16,13
topic : 415	6	54	11,11	6	12	50,00
topic : 416	12	23	52,17	12	31	38,71
topic : 417	1	1	100,00	1	4	25,00
topic : 418	1	14	7,14	1	10	10,00
topic : 419	37	67	55,22	37	56	66,07
topic : 420	18	34	52,94	18	30	60,00
topic : 421	18	65	27,69	18	37	48,65
topic : 422	9	15	60,00	9	54	16,67
topic : 423	35	56	62,50	35	54	64,81

topic : 424	184	199	92,46	184	196	93,88
topic : 425	179	294	60,88	179	212	84,43
topic : 426	24	74	32,43	24	54	44,44
topic : 427	16	63	25,40	16	55	29,09
topic : 428	107	154	69,48	107	164	65,24
topic : 429	76	227	33,48	76	117	64,96
topic : 430	113	141	80,14	113	161	70,19
topic : 431	63	125	50,40	63	105	60,00
topic : 433	9	15	60,00	9	23	39,13
topic : 434	79	109	72,48	79	91	86,81
topic : 435	41	53	77,36	41	46	89,13
topic : 436	1	1	100,00	1	5	20,00
topic : 439	0	0	#DIV/0!	0	1	0,00
topic : 440	91	124	73,39	91	133	68,42
topic : 441	46	79	58,23	46	70	65,71
topic : 444	0	6	0,00	0	11	0,00
topic : 445	118	181	65,19	118	147	80,27
topic : 446	8	51	15,69	8	64	12,50
topic : 447	35	68	51,47	35	44	79,55
topic : 448	135	172	78,49	135	167	80,84
topic : 449	118	133	88,72	118	122	96,72
topic : 450	3	5	60,00	3	6	50,00
topic : 453	1	3	33,33	1	4	25,00
topic : 454	79	114	69,30	79	122	64,75
topic : 458	152	198	76,77	152	184	82,61
topic : 459	49	72	68,06	49	82	59,76
topic : 461	2	2	100,00	2	2	100,00
topic : 463	17	58	29,31	17	48	35,42
topic : 464	111	138	80,43	111	122	90,98
topic : 465	57	136	41,91	57	97	58,76
topic : 467	1	2	50,00	1	5	20,00
topic : 468	0	2	0,00	0	4	0,00
topic : 469	47	75	62,67	47	76	61,84
topic : 470	24	44	54,55	24	43	55,81
topic : 471	7	10	70,00	7	28	25,00
topic : 472	15	40	37,50	15	76	19,74
topic : 473	14	31	45,16	14	24	58,33
topic : 474	39	90	43,33	39	87	44,83
topic : 475	51	115	44,35	51	126	40,48
topic : 476	43	59	72,88	43	49	87,76
topic : 477	29	47	61,70	29	43	67,44
topic : 478	7	14	50,00	7	12	58,33
topic : 479	20	55	36,36	20	53	37,74
topic : 480	14	27	51,85	14	25	56,00
topic : 481	3	13	23,08	3	20	15,00

topic : 482	183	201	91,04	183	202	90,59
topic : 483	76	145	52,41	76	142	53,52
topic : 484	3	10	30,00	3	3	100,00
topic : 485	53	81	65,43	53	70	75,71
topic : 486	17	34	50,00	17	30	56,67
topic : 487	10	25	40,00	10	38	26,32
topic : 488	81	170	47,65	81	109	74,31
topic : 489	39	95	41,05	39	90	43,33
topic : 490	35	75	46,67	35	50	70,00
topic : 491	1	5	20,00	1	5	20,00
topic : 495	5	26	19,23	5	33	15,15
topic : 496	27	44	61,36	27	43	62,79
topic : 497	27	88	30,68	27	58	46,55
topic : 498	9	75	12,00	9	21	42,86
topic : 499	13	13	100,00	13	13	100,00
topic : 500	10	46	21,74	10	36	27,78
topic : 502	166	216	76,85	166	187	88,77
topic : 503	22	81	27,16	22	63	34,92
topic : 505	66	140	47,14	66	148	44,59
topic : 506	62	93	66,67	62	93	66,67
topic : 507	13	13	100,00	13	23	56,52
topic : 508	11	54	20,37	11	49	22,45
topic : 509	219	219	100,00	219	219	100,00
topic : 511	0	0	0	0	0	0
topic : 515	6	23	26,09	6	28	21,43
topic : 516	8	31	25,81	8	10	80,00
topic : 517	1	16	6,25	1	8	12,50
topic : 518	78	155	50,32	78	148	52,70
topic : 519	39	47	82,98	39	45	86,67
topic : 520	36	57	63,16	36	50	72,00
topic : 521	21	56	37,50	21	43	48,84
topic : 522	11	31	35,48	11	32	34,38
topic : 523	38	69	55,07	38	91	41,76
topic : 525	19	51	37,25	19	31	61,29
topic : 526	81	229	35,37	81	149	54,36
topic : 527	57	174	32,76	57	118	48,31
topic : 528	210	221	95,02	210	218	96,33
topic : 529	21	72	29,17	21	33	63,64
topic : 530	97	114	85,09	97	111	87,39
topic : 531	10	62	16,13	10	21	47,62
topic : 532	1	30	3,33	1	15	6,67
topic : 533	44	95	46,32	44	64	68,75
topic : 534	74	106	69,81	74	99	74,75
topic : 535	17	24	70,83	17	20	85,00
topic : 536	64	106	60,38	64	89	71,91

topic : 537	16	66	24,24	16	28	57,14
topic : 538	1	5	20,00	1	15	6,67
topic : 539	14	25	56,00	14	17	82,35
topic : 540	0	2	0,00	0	0	#DIV/0!
topic : 541	16	44	36,36	16	31	51,61
topic : 542	20	91	21,98	20	48	41,67
topic : 543	11	61	18,03	11	27	40,74
Total	4529	8025	56,44	4529	6959	65,08

Table B.3. Percentage of link relevance (107 CO topics of INEX 2007, with top 100 relevant documents retrieved)

	IN			OUT		
TOPIC_ID	Relevant links	All links	% relevant	Relevant links	All links	% relevant
topic : 414	10	139	7,19	10	49	20,41
topic : 415	10	107	9,35	10	32	31,25
topic : 416	17	38	44,74	17	58	29,31
topic : 417	1	1	100,00	1	8	12,50
topic : 418	1	15	6,67	1	11	9,09
topic : 419	37	73	50,68	37	56	66,07
topic : 420	21	54	38,89	21	39	53,85
topic : 421	28	126	22,22	28	66	42,42
topic : 422	20	42	47,62	20	90	22,22
topic : 423	69	125	55,20	69	109	63,30
topic : 424	410	493	83,16	410	452	90,71
topic : 425	212	400	53,00	212	263	80,61
topic : 426	34	128	26,56	34	101	33,66
topic : 427	34	166	20,48	34	125	27,20
topic : 428	160	291	54,98	160	303	52,81
topic : 429	142	407	34,89	142	201	70,65
topic : 430	134	175	76,57	134	204	65,69
topic : 431	98	207	47,34	98	162	60,49
topic : 433	15	26	57,69	15	66	22,73
topic : 434	100	160	62,50	100	120	83,33
topic : 435	77	108	71,30	77	95	81,05
topic : 436	1	1	100,00	1	7	14,29
topic : 439	1	2	50,00	1	5	20,00
topic : 440	125	208	60,10	125	208	60,10
topic : 441	80	153	52,29	80	140	57,14
topic : 444	1	14	7,14	1	19	5,26
topic : 445	202	338	59,76	202	261	77,39
topic : 446	8	83	9,64	8	109	7,34
topic : 447	42	126	33,33	42	62	67,74

topic : 448	194	303	64,03	194	250	77,60
topic : 449	189	245	77,14	189	205	92,20
topic : 450	6	10	60,00	6	9	66,67
topic : 453	2	9	22,22	2	8	25,00
topic : 454	93	158	58,86	93	160	58,13
topic : 458	191	333	57,36	191	279	68,46
topic : 459	60	104	57,69	60	116	51,72
topic : 461	2	3	66,67	2	3	66,67
topic : 463	18	93	19,35	18	72	25,00
topic : 464	324	454	71,37	324	399	81,20
topic : 465	77	198	38,89	77	139	55,40
topic : 467	1	2	50,00	1	5	20,00
topic : 468	0	2	0,00	0	4	0,00
topic : 469	65	132	49,24	65	120	54,17
topic : 470	25	60	41,67	25	54	46,30
topic : 471	7	10	70,00	7	39	17,95
topic : 472	15	55	27,27	15	116	12,93
topic : 473	14	35	40,00	14	24	58,33
topic : 474	62	201	30,85	62	177	35,03
topic : 475	53	155	34,19	53	154	34,42
topic : 476	61	79	77,22	61	68	89,71
topic : 477	35	62	56,45	35	56	62,50
topic : 478	27	45	60,00	27	39	69,23
topic : 479	44	105	41,90	44	110	40,00
topic : 480	26	54	48,15	26	47	55,32
topic : 481	3	19	15,79	3	26	11,54
topic : 482	247	318	77,67	247	337	73,29
topic : 483	85	199	42,71	85	202	42,08
topic : 484	3	15	20,00	3	7	42,86
topic : 485	57	106	53,77	57	90	63,33
topic : 486	17	36	47,22	17	31	54,84
topic : 487	16	84	19,05	16	61	26,23
topic : 488	151	413	36,56	151	221	68,33
topic : 489	94	247	38,06	94	187	50,27
topic : 490	35	101	34,65	35	55	63,64
topic : 491	12	23	52,17	12	23	52,17
topic : 495	5	38	13,16	5	43	11,63
topic : 496	38	74	51,35	38	65	58,46
topic : 497	50	138	36,23	50	95	52,63
topic : 498	9	129	6,98	9	46	19,57
topic : 499	222	232	95,69	222	232	95,69
topic : 500	14	77	18,18	14	48	29,17
topic : 502	195	306	63,73	195	262	74,43
topic : 503	29	106	27,36	29	85	34,12
topic : 505	112	282	39,72	112	245	45,71

topic : 506	94	144	65,28	94	155	60,65
topic : 507	37	37	100,00	37	63	58,73
topic : 508	11	96	11,46	11	66	16,67
topic : 509	392	422	92,89	392	408	96,08
topic : 511	0	0	0	0	0	0
topic : 515	15	38	39,47	15	44	34,09
topic : 516	12	50	24,00	12	18	66,67
topic : 517	1	34	2,94	1	12	8,33
topic : 518	103	231	44,59	103	205	50,24
topic : 519	49	67	73,13	49	66	74,24
topic : 520	49	108	45,37	49	82	59,76
topic : 521	24	86	27,91	24	68	35,29
topic : 522	11	48	22,92	11	46	23,91
topic : 523	59	104	56,73	59	138	42,75
topic : 525	19	72	26,39	19	39	48,72
topic : 526	86	329	26,14	86	176	48,86
topic : 527	116	360	32,22	116	247	46,96
topic : 528	419	448	93,53	419	443	94,58
topic : 529	33	126	26,19	33	56	58,93
topic : 530	106	143	74,13	106	128	82,81
topic : 531	15	114	13,16	15	38	39,47
topic : 532	10	67	14,93	10	42	23,81
topic : 533	108	214	50,47	108	167	64,67
topic : 534	79	157	50,32	79	113	69,91
topic : 535	30	45	66,67	30	37	81,08
topic : 536	108	210	51,43	108	157	68,79
topic : 537	23	94	24,47	23	41	56,10
topic : 538	2	9	22,22	2	25	8,00
topic : 539	26	50	52,00	26	35	74,29
topic : 540	0	9	0,00	0	7	0,00
topic : 541	22	84	26,19	22	57	38,60
topic : 542	20	160	12,50	20	81	24,69
topic : 543	11	95	11,58	11	44	25,00
Total	7165	14307	50,08	7165	11739	61,04