Evaluation of Effective XML Information Retrieval

A thesis submitted for the degree of
Doctor of Philosophy

Jovan Pehcevski, Grad. Electro-Technical Eng.
(University "St. Cyril and Methodius", Skopje, Republic of Macedonia)
School of Computer Science and Information Technology,
Science, Engineering, and Technology Portfolio,
RMIT University,
Melbourne, Victoria, Australia.

28th August, 2006

This thesis is dedicated to my father, Dimitar Pehcevski,
who never stopped believing

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Jovan Pehcevski
School of Computer Science and Information Technology
RMIT University
28th August, 2006

Acknowledgments

I've come up with a set of rules that describe our reactions to technologies:

1. *Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works.*
2. *Anything that's invented between when you're fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it.*
3. *Anything invented after you're thirty-five is against the natural order of things.*

— Douglas Adams, *The Salmon of Doubt*, Harmony Books, 2002.

There is a common wisdom that every saga has a beginning. There is another not so common wisdom that the road towards a PhD degree does not always have an end (or at least not a satisfactory one). I believe I am one of the few blessed people in this world to have a companion like Irena, my real life partner and my immortal beloved, in travelling the road less travelled. I thank her for tolerating my (at times) peculiar research behaviour, my weird passion towards science-fiction books and movies, all my sleepless nights, and for everything that has happened in between during the last four crazy years. I do hope that, one bright and sunny day, I will be able to make all of this up to her. I also want to express my deepest gratitudes to members of our closest families: my mother Jordanka, my brother Dragan and his lovely wife Lidija, my grandfather Jovan and my other two grandparents, my uncle Venko with his family, Irena's father Metodi, Irena's mother Jelica, and Irena's brother Vlado. I thank them for their never-ending love and support, and for believing in our goals.

On my academic side, my deepest gratitudes go to my first thesis adviser, Dr. James Thom, for his thorough guidance during my PhD candidature. It was Jamie who introduced me to the academic research, who opened the doors to its endless opportunities, and who to this day is probably still wondering why I was writing papers while running out of time to write my thesis. His research experience and continuous feedback were invaluable for the achieved quality of my research work. I thank Dr. Seyed Mohammad Mehdi (Saied) Tahaghoghi, my second thesis adviser and my favourite lecturer, for his many useful writing suggestions and technical tips, and for his generous academic and personal support. Saied introduced me to the world of academic teaching, and assisted me to learn a great deal about

effective teaching techniques. I am also indebted to Dr. Anne-Marie Vercoustre, my valuable consultant and a former second thesis adviser, for her encouragement towards achieving my PhD goal. I am personally looking forward to continue our collaboration with Anne-Marie at INRIA later this year. I also want to express my gratitudes to Michael Harris and Donal Ellis, two great Aussie mates, who taught me many useful course management tips and techniques. Finally, I thank Prof. Justin Zobel for managing our vibrant research group, and for organising the "Theory of IR" meetings where I learned a lot about the activities surrounding the exciting field of information retrieval.

On my socialising side, I thank Tome Jovanovski and Jasmina Karevski for their honest friendship during the last three years. They shared many of the joyful and most memorable moments with us here in Australia, and for that Irena and I are most grateful. There are some past and present members of my RMIT research group that I especially want to thank: Timo Volkmer, for opening the door into the world of Douglas Adams; Abhijit Chattaraj, for discussing and sharing with me the intricacies of the world of Douglas Adams; Falk Scholer, for our joint admiration of the world of elves, dwarfs, and wizards, which does not have anything to do with the world of Douglas Adams; Martin Plowman, for reminding me that I could well be one of the legal aliens that he is so desperately seeking to interview for his PhD thesis; Vaughan Shanks, for his many live demonstrations of the great Aussie spirit; and Ranjan Sinha, Bodo Billerbeck, Halil Ali, and Steven Garcia, for just being there.

Last, but definitely not least, I want to thank some of the academics being part of the wonderful INEX research community; indeed, without INEX, the research work presented in this thesis would not have been possible. I thank Prof. Mounia Lalmas, for her ongoing encouragement and support; Andrew Trotman, for his friendship and for our very many useful research discussions; Shlomo Geva, for his useful insights and always intriguing e-mails; and Benjamin Piwowarski, Birger Larsen, Tassos Tombros, Jaap Kamps, Gabriella Kazai, Arjen de Vries, Djoerd Hiemstra, and Miro Lehtonen, for engaging with me into many interesting and helpful e-mail discussions. I sincerely thank all of them for their help and support, and I truly hope we will continue our successful collaboration in future.

In the meantime, I continue to be very excited and enthusiastic about all things concerning *life*, the *universe*, and *everything*. But that's most likely because I am still only thirty-four.

— Jovan Pehcevski, Melbourne, Australia, August 2006

Credits

Portions of the material in this thesis have previously appeared in the following publications:

- "Hybrid XML Retrieval: Combining Information Retrieval and a Native XML Database", *Information Retrieval* [Pehcevski et al., 2005b] (Chapter 3).
- "Relevance in XML Retrieval: The User Perspective", *Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology* [Pehcevski, 2006] (Chapter 4).
- "HiXEval: Highlighting XML Retrieval Evaluation", *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), LNCS 3977* [Pehcevski and Thom, 2006] (Chapters 4 and 5).
- "RMIT University at INEX 2005: Ad hoc Track", *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), LNCS 3977* [Pehcevski et al., 2006] (Chapter 6).
- "Combining Image and Structured Text Retrieval", *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), LNCS 3977* [Awang Iskandar et al., 2006] (Chapter 6 and Appendix C).
- "Users and Assessors in the Context of INEX: Are Relevance Dimensions Relevant?", *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology* [Pehcevski et al., 2005c] (Chapter 4 and Appendix B).
- "Hybrid XML Retrieval Revisited", *Advances in XML Information Retrieval: Third International Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2004), LNCS 3493* [Pehcevski et al., 2005a] (Chapter 3).
- "Enhancing Content-And-Structure Information Retrieval using a Native XML Database", *Proceedings of the first Twente Data Management Workshop (TDM'04) on XML Databases and Information Retrieval* [Pehcevski et al., 2004b] (Chapter 3).
- "RMIT INEX Experiments: XML Retrieval using Lucy/eXist", *Proceedings of the INEX 2003 workshop* [Pehcevski et al., 2004a] (Chapter 3).

- "XML-Search Query Language: Needs and Requirements", *Proceedings of the AUSWeb 2003 conference* [Pehcevski et al., 2003] (Chapters 1 and 2).

This work was supported by the International Postgraduate Research Scholarship (IPRS), the Australian Research Council and the Search Engine Group at RMIT University.
The thesis was written using the `KWrite` editor on Mandrake GNU/Linux, and typeset using the LaTeX 2_ε document preparation system.

Note

Unless otherwise stated, all fractional results have been rounded to the displayed number of decimal figures.

Contents

Glossary of acronyms

A-overlap	Ascendants overlap
AP	Average Precision
BA	Broad Answer
BEP	Best Entry Point
CAS	Content And Structure
CBEP	Combined BEP
CBIR	Content-Based Image Retrieval
CG	Cumulated Gain
CO	Content Only
CO+S	Content Only plus Structure
CRE	Coherent Retrieval Element
D-overlap	Descendants overlap
DFR	Divergence From Randomness
EA	Exact Answer
ep/gr	effort precision / gain recall
EPRUM	Expected Precision-Recall with User Modelling
Ex	Exhaustivity
F@r	F-measure (harmonic mean between Precision and Recall) at rank r
FS	Fully Seen
FullRB	Full Recall-Base
GIFT	GNU Image Finding Tool
HSV	Hue Saturation Value
HiXEval	Highlighting XML Retrieval Evaluation
HyREX	Hypermedia Retrieval Engine for XML

IDF	Inverse Document Frequency
IDL	Inverse Document Length
IEF	Inverse Element Frequency
iMAP	interpolated Mean Average Precision
INEX	INitiative for the Evaluation of XML Retrieval
IR	Information Retrieval
LCA	Lowest Common Ancestor
MA	Mutually Agreed
MAP	Mean Average Precision
MAep	Mean Average effort precision
MM	Multimedia
MpE	Ranking heuristic: more matching elements (M), shorter XPath length (p), nearer to end (E)
NA	Narrow Answer
nCG	normalised Cumulated Gain
nCRE	CRE that represents either LCA of at least two matching elements, or a matching element whose parent is not recognised as LCA
NEXI	Narrowed Extended XPath I
NR	Not Relevant
NS	Not Seen
nxCG	normalised extended Cumulated Gain
oCRE	CRE that represents LCA of at least two matching elements
O-overlap	Overall overlap
P-overlap	Probabilistic overlap
P@r	Precision at rank r
PA	Partial Answer
PBEP	Parent BEP
PME	Ranking heuristic: longer XPath length (P), more matching elements (M), nearer to end (E)
PS	Partially Seen
PTF	Ranking heuristic: longer XPath length (P), more distinct query terms (T), more frequent query term occurrences (F)
Prec	Precision

R-prec	Mean Recall-precision
R@r	Recall at rank r
RMIT	Royal Melbourne Institute of Technology
RP	Recall-precision
RSV	Retrieval Status Value
Rec	Recall
SBEP	Start-reading-here BEP
SCAS	Strict Content And Structure
SS	CAS query interpretation: strict target and strict support elements
SV	CAS query interpretation: strict target and vague support elements
Sp	Specificity
T2I	Tolerance to Irrelevance
TF	Term Frequency
TPF	Ranking heuristic: more distinct query terms (T), longer XPath length (P), more frequent query term occurrences (F)
TREC	Text REtrieval Conference
Trel	Total amount of relevant information
VCAS	Vague Content And Structure
VS	CAS query interpretation: vague target and strict support elements
VV	CAS query interpretation: vague target and vague support elements
WWW	World Wide Web
W3C	World Wide Web Consortium
XCG	Extended Cumulated Gain
XML	eXtensible Markup Language
XPath	XML Path language
XQuery	XML Query language
XSL	eXtensible Stylesheet Language

List of Figures

List of Tables

Abstract

XML is being adopted as a common storage format in scientific data repositories, digital libraries, and on the World Wide Web. Accordingly, there is a need for content-oriented XML retrieval systems that can efficiently and effectively store, search and retrieve information from XML document collections. Unlike traditional information retrieval systems where whole documents are usually indexed and retrieved as information units, XML retrieval systems typically index and retrieve document components of varying granularity. To evaluate the effectiveness of such systems, test collections — where relevance assessments are provided according to an XML-specific definition of relevance — are necessary. Such test collections have been built during four rounds of the INitiative for the Evaluation of XML Retrieval (INEX).

There are many different approaches to XML retrieval; most approaches either extend full-text information retrieval systems to handle XML retrieval, or use database technologies that incorporate existing XML standards to handle both XML presentation and retrieval. We present a *hybrid approach* to XML retrieval that combines text information retrieval features with XML-specific features found in a native XML database. Results from our experiments on the INEX 2003 and 2004 test collections demonstrate the usefulness of applying our hybrid approach to different XML retrieval tasks.

A realistic definition of *relevance* is necessary for meaningful comparison of alternative XML retrieval approaches. The three relevance definitions used by INEX since 2002 comprise two relevance dimensions, each based on topical relevance. We perform an extensive analysis of the two INEX 2004 and 2005 relevance definitions, and show that assessors and users find them difficult to understand. We propose a new definition of relevance for XML retrieval, and demonstrate that a relevance scale based on this definition is useful for XML retrieval experiments.

Finding the appropriate approach to evaluate XML retrieval effectiveness is the subject of ongoing debate within the XML information retrieval research community. We present an overview of the evaluation methodologies implemented in the current INEX metrics, which reveals that the metrics follow different assumptions and measure different XML retrieval behaviours. We propose a new *evaluation metric* for XML retrieval and conduct an extensive analysis of the retrieval performance of simulated runs to show what is measured. We compare the evaluation behaviour obtained with the new metric to the behaviours obtained with two of the official INEX 2005 metrics, and demonstrate that the new metric can be used to reliably evaluate XML retrieval effectiveness.

To analyse the effectiveness of XML retrieval in different *application scenarios*, we use evaluation measures in our new metric to investigate the behaviour of XML retrieval approaches under the following two scenarios: the ad-hoc retrieval scenario, exploring the activities carried out as part of the INEX 2005 Ad-hoc track; and the multimedia retrieval scenario, exploring the activities carried out as part of the INEX 2005 Multimedia track. For both application scenarios we show that, although different values for retrieval parameters are needed to achieve the optimal performance, the desired textual or multimedia information can be effectively located using a combination of XML retrieval approaches.