

OSWIR 2005

First workshop on

Open Source Web Information Retrieval

Edited by Michel Beigbeder and Wai Gen Yee

ISBN: 2-913923-19-4

OSWIR 2005 Organization

Workshop chairs

Michel Beigbeder
G2I department
École Nationale Supérieure des Mines de Saint-Étienne, France

Wai Gen Yee
Department of Computer Science
Illinois Institute of Technology, USA

Program Committee

Abdur Chowdhury, America Online Search and Navigation, USA
Ophir Frieder, Illinois Institute of Technology, USA
David Grossman, Illinois Institute of Technology, USA
Donald Kraft, Louisiana State University, USA
Clement Yu, University of Illinois at Chicago, USA

Reviewers

Jefferson Heard, Illinois Institute of Technology, USA
Dongmei Jia, Illinois Institute of Technology, USA
Linh Thai Nguyen, Illinois Institute of Technology, USA

OSWIR 2005

Open Source Web Information Retrieval

The World Wide Web has grown to be a primary source of information for millions of people. Due to the size of the Web, search engines have become the major access point for this information. However, "commercial" search engines use hidden algorithms that put the integrity of their results in doubt, so there is a need for some open source Web search engines.

On the other hand, the Information Retrieval (IR) research community has a long history of developing ideas, models and techniques for finding results in data sources, but finding one's way through all of them is not an easy task. Moreover their applicability to the Web search domain is uncertain.

The goal of the workshop is to survey the fundamentals of the IR domain and to determine the techniques, tools, or models that are applicable to Web search.

This first workshop was organized by Michel BEIGBEDER from *École Nationale Supérieure des Mines de Saint-Étienne*¹, France and Wai Gen YEE from *Illinois Institute of Technology*², USA. It was held on September 19th, 2005 in the UTC (*Compiègne University of Technology*)³ in conjunction with WI and IAT 2005⁴ the 2005 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

We want to thank all the authors of the submitted papers, the members of the program committee: Abdur Chowdhury, Ophir Frieder, David Grossman, Donald Kraft, Clement Yu and the reviewers: Jefferson Heard, Dongmei Jia, and Linh Thai Nguyen.

Michel Beigbeder and Wai Gen Yee

¹<http://www.emse.fr/>

²<http://www.iit.edu/>

³<http://www.hds.utc.fr/>

⁴<http://www.hds.utc.fr/WI05/>

Table of contents

Pre-processing Text for Web Information Retrieval Purposes by Splitting Compounds into their Morphemes <i>Sven Abels and Axel Hahn</i>	7
Fuzzy Querying of XML documents – The Minimum Spanning Tree <i>Abdeslame Alilaouar and Florence Sedes</i>	11
Link Analysis in National Web Domains <i>Ricardo Baeza-Yates and Carlos Castillo</i>	15
Web Document Models for Web Information Retrieval <i>Michel Beigbeder</i>	19
Static Ranking of Web Pages, and Related Ideas <i>Wray Buntine</i>	23
WIRE: an Open Source Web Information Retrieval Environment <i>Carlos Castillo and Ricardo Baeza-Yates</i>	27
Nutch: an Open-Source Platform for Web Search <i>Doug Cutting</i>	31
Towards Contextual and Structural Relevance Feedback in XML Retrieval <i>Lobna Hlaoua and Mohand Boughanem</i>	35
An Extension to the Vector Model for Retrieving XML Documents <i>Fabien Laniel and Jean-Jacques Girardot</i>	39
Do Search Engines Understand Greek or User Requests "Sound Greek" to them? <i>Fotis Lazarinis</i>	43
Use of Kolmogorov Distance Identification of Web Page Authorship, Topic and Domain <i>David Parry</i>	47
Searching Web Archive Collections <i>Michael Stack</i>	51
XGTagger, an Open-Source Interface Dealing with XML Contents <i>Xavier Tannier, Jean-Jacques Girardot and Mihaela Mathieu</i>	55
The Lifespan, Accessibility and Archiving of Dynamic Documents <i>Katarzyna Wegrzyn-Wolska</i>	59
SYRANNOT: Information Retrieval Assistance System on the Web by Semantic Annotations Re-use <i>Wiem Yaiche Elleuch, Lobna Jeribi and Abdelmajid Ben Hamadou</i>	63
Search in Peer-to-Peer File-Sharing System: Like Metasearch Engines, But Not Really <i>Wai Gen Yee, Dongmei Jia and Linh Thai Nguyen</i>	67

