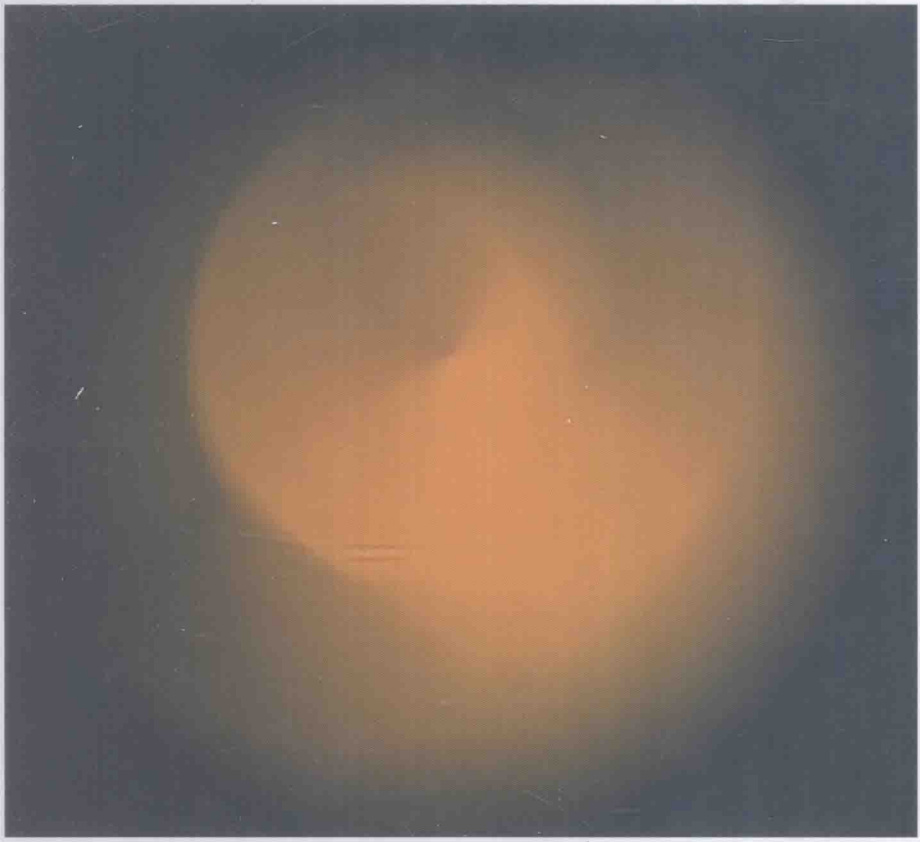
Intelligent Information Systems – Vol. 3

REASONING ABOUT FUZZY TEMPORAL AND SPATIAL INFORMATION FROM THE WEB

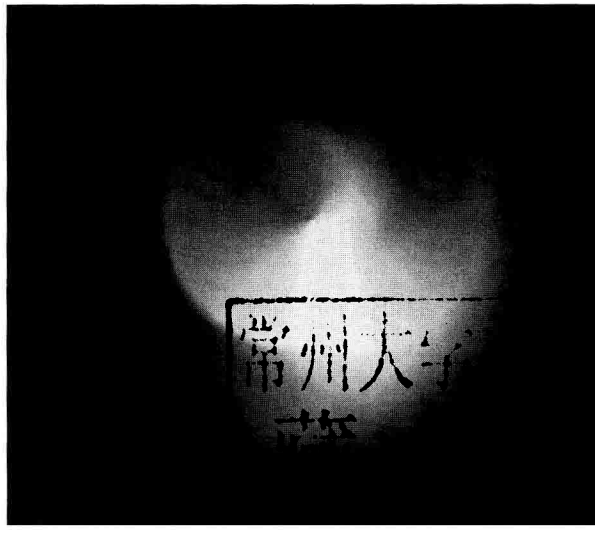


Steven Schockaert • Martine De Cock • Etienne Kerre

World Scientific

Intelligent Information Systems – Vol. 3

REASONING ABOUT FUZZY TEMPORAL AND SPATIAL INFORMATION FROM THE WEB



Steven Schockaert • Martine De Cock • Etienne Kerre
Ghent University, Belgium

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

REASONING ABOUT FUZZY TEMPORAL AND SPATIAL INFORMATION FROM THE WEB

Intelligent Information Systems — Vol. 3

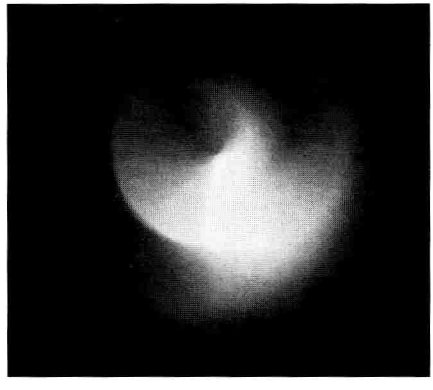
ISBN-13 978-981-4307-89-5

ISBN-10 981-4307-89-0

Printed in Singapore by World Scientific Printers.

Intelligent Information Systems – Vol. 3

REASONING ABOUT FUZZY TEMPORAL AND SPATIAL INFORMATION FROM THE WEB



INTELLIGENT INFORMATION SYSTEMS

Series Editors: Da Ruan *(Belgian Nuclear Research Centre (SCK.CEN) & Ghent University, Belgium)*
Jie Lu *(University of Technology, Sdyney, Australia)*

Vol. 1 Advances in Artificial Intelligence for Privacy Protection and Security
edited by Agusti Solanas & Antoni Martínez-Ballesté (Rovira i Virgili University, Spain)

Vol. 2 E-Business in the 21st Century: Realities, Challenges and Outlook
by Jun Xu (Southern Cross University, Australia) & Mohammed Quaddus (Curtin University of Technology, Australia)

Vol. 3 Reasoning about Fuzzy Temporal and Spatial Information from the Web
by Steven Schockaert (Ghent University, Belgium), Martine De Cock (Ghent University, Belgium) & Etienne Kerre (Ghent University, Belgium)

Preface

Traditionally, research in artificial intelligence (AI) has centered around formal representations of information, with the main focus lying on the expressivity of different knowledge representation languages, their computational complexity, and the development of sound and complete reasoners. In information retrieval (IR), on the other hand, the key notion has been manipulating numbers of occurrences of terms, for instance to find statistical evidence that a given text document is relevant to a given textual query, and the most important focus has been on scalability, robustness and the precision-recall trade-off of different retrieval models. Due to a significant progress in domains such as information extraction and natural language processing, however, hybrid approaches have recently emerged, which try to apply symbolic reasoning on information that has been automatically extracted from natural language text. This observation is related to the more general trend of trying to endow search engines with more intelligence by restricting search engine results to what is actually needed to meet the user's information need (focused information access), rather than simply providing a list of possibly relevant documents.

Symbolic reasoning in IR has among others been studied for the analysis of news stories and for building question answering systems, two areas in which temporal information is prevalent. Indeed, temporal relations and events appear to be fundamental both in natural language processing and artificial intelligence, and if automated reasoning from natural language is to be successful, a thorough understanding of the temporal structure of the underlying discourse is required. When moving from news stories to information from the web, however, traditional symbolic reasoning seems not applicable anymore, due to the informal nature of the language, and of the topics that are discussed. For example, while most people would

think of the Cold War as something which has happened after the Second World War, many historians date the beginning of the Cold War as early as the end of the 1917 Russian Revolution. The existence of such types of disagreement on the temporal boundaries of events calls for extensions of classical formalisms

Disagreement on the boundaries of categories is most naturally modeled using fuzzy set theory, and various models for representing fuzzy event boundaries have already been proposed. In this existing work, however, only the modeling aspect has been studied, and reasoning with fuzzy temporal information has hardly been covered at all. Thus there seems to be a gap between what has been done on temporal reasoning in AI, focusing on efficiency and reasoning, and what has been done in the context of fuzzy set theory, focusing on expressivity and modeling. It is precisely this missing link between AI and fuzzy set theory that seems crucial to allow reasoning about the kind of informal information we find on the web.

This book is an attempt to find this missing link in the temporal and spatial domains. After three introductory chapters, also introducing the required technical background, fuzzy temporal information processing is studied from three different angles in Chapters 4–6: modeling fuzzy temporal information in Chapter 4 (fuzzy set theory), reasoning with fuzzy temporal information in chapter 5 (AI) and applying fuzzy temporal information on web information in Chapter 6 (IR). As such, our focus is both on obtaining new theoretical insights, and on applying and evaluating them in a practical setting. Subsequently, we also look at fuzzy spatial information from the same three angles in Chapter 7–9. In studying both temporal and spatial information, we discover many parallels, but at the same time, we also lay bare a number of surprising differences between how fuzzy temporal and spatial information behaves.

This book originated from the doctoral thesis of the first author, which was successfully defended in April 2008. Encouraged by the enthusiastic reports of the committee members, and by two awards that have been given to this thesis (the ECCAI Artificial Intelligence Dissertation Award, and the FWO/IBM Belgium Prize for Computer Science), we have decided to publish this book, and make the obtained results available to a larger audience. We are grateful to Alia Abdelmoty, David Ahn, Chris Cornelis, Maarten de Rijke, Christopher Jones, Philip Smart, and Florian Twaroch for many fruitful discussions, and their comments and suggestions which have clearly influenced the results in this work. We also would like to thank the external members of the reading committee, Anthony Cohn,

Didier Dubois, and Dirk Vermeir, for their useful suggestions on the first version of the thesis, and Da Ruan for his help with the publication of this book. Finally, we would like to thank the Research Foundation – Flanders (FWO) for the financial support.

Contents

Chapter 1

Introduction

Although the central focus of this book is not on information retrieval (IR), it is in this domain that the main motivation for our work is rooted. In this introductory chapter, we provide a glimpse at the field of IR, highlighting how research is increasingly moving towards more "intelligent" techniques and a more thorough use of semantics. In the first section, we focus on the classical paradigm of document retrieval, which has initially been developed in the 1970s and 1980s, but received a tremendous boost in the 1990s with the advent of the web. The next section sketches a number of more advanced information access paradigms, adopting linguistic processing as the main vehicle for achieving intelligence. Subsequently, in Section 1.3, a number of recent trends are discussed, focusing on semantic approaches for retrieving objects, rather than documents. In a final section, we argue that the tendency towards more semantics in IR goes hand in hand with an increased need for appropriate models of time and space. We furthermore emphasize the informal nature of available temporal and spatial information and the resulting need to explicitly deal with vagueness. To conclude, we provide an overview of how these issues will be addressed in the remainder of this book.

1.1 Document Retrieval

The field IR, in general, is concerned with assisting users in acquiring information of interest. By far the most dominant IR paradigm is based on users formulating keyword queries such as *Norway hiking national parks* to express information needs (e.g., Would Norway be an appropriate holiday destination for me?). These queries are subsequently used by the system to estimate the relevance of each document in the collection of interest.

Finally, the most relevant documents are presented to the user in the form of a ranked list. Although an abundance of mathematical models exists for estimating the degree of relevance of a document, they are virtually always based on the same two principles:

(1) the higher the number of occurrences of the query terms in a document, the more likely it is relevant;
(2) the fewer documents a particular query term appears in, the more weight should be given to it.

Thus, documents are essentially reduced to bags of words, as calculations only depend on the number of times a given term appears in a given document. Note, however, that this basic model has been extended along various lines and state–of–the–art systems additionally incorporate features such as the proximity of query terms in a document, allow the use of phrases in queries, etc. Nonetheless, the techniques employed seem surprisingly simple: documents are selected and ranked without any attempt to grasp the semantics of either the query or the documents.

A wide array of more "intelligent" techniques have been proposed to improve the performance of document retrieval systems, albeit with mixed success. A recurring theme is the semantic gap between the query terms and the actual terms used in the document. In particular, most words in English, and many other languages, can have different meanings in different contexts (polysemy), and different words can still have the same meaning (synonymy). A significant increase in performance can therefore be expected when document and query terms are mapped to unambiguous concepts (word sense disambiguation). As different synonyms are mapped to the same concept, more relevant documents should be found. Moreover, as polysemous words are mapped to different concepts in different contexts, less irrelevant documents should be returned: only if the polysemous word is used in the same sense in both query and document, the document will be considered relevant. Experimental results along these lines have been largely disappointing, however [Sanderson (2000); Smeaton (1999); Voorhees (1999)]. The main conclusion drawn in [Voorhees (1999)] is that linguistic techniques have to be essentially perfect to be helpful, which they are not. For example, if the algorithm for disambiguating polysemous words is too error–prone, more relevant documents are missed because of these errors than discovered because of using concepts. In particular, it turns out to be extremely difficult to disambiguate the query terms, as only very little context is available to this end, viz., the other query terms. Along similar

lines, in [Mihalcea and Moldovan (2001)] it is proposed to apply a named entity (NE) recognizer to recognize entities such as persons, locations and organizations in documents. This information could help to locate relevant documents if it is known, in advance, that the desired information is concerned with, e.g., a person. Similar considerations as for word sense disambiguation apply. For example, in [Chu-Carroll and Prager (2007)], an experimental study is conducted to analyse the relationship between the accuracy of the NE recognizer and the effect on retrieval performance, i.e., how good an NE recognizer should be in order to be helpful. Other natural language processing (NLP) techniques that have been applied to IR are part–of–speech tagging (e.g., [Kraaij and Pohlmann (1996)]), following an assumption that nouns in documents should influence the degree of relevance more than verbs or adjectives, and noun phrase chunking (e.g., [Evans and Zhai (1996)]), following an assumption that linguistically motivated phrases are better suited for estimating a document's relevance score than statistically motivated phrases. Again, experimental results are mixed, revealing positive effects in some cases and negative effects in others. This seems to suggest that NLP is not as paramount in information retrieval as was originally thought. Note, however, that some simple linguistically oriented techniques are nevertheless fundamental in current IR systems, in particular the removal of stop words (i.e., non–content words such as "and", "he", "are", etc.) and stemming (i.e., removing certain suffices from words). The use of more advanced techniques seems to be impeded by the current state–of–the–art in NLP. Another generally accepted explanation for the disappointing results of NLP techniques is the observation that the statistical techniques utilized in IR implicitly capture more linguistic meaning than intuition would suggest.

A strategy which is closely related to NLP is the use of machine readable dictionaries, or thesauri, to bridge the semantic gap. The main hypothesis here, as for word sense disambiguation, is that many relevant documents are missed because the query terms are not exactly the same as the terms that occur in some relevant documents. Documents might use synonyms, but also hypernyms (i.e., more general terms) or hyponyms (i.e., more specific terms) of the query terms. In particular, it is assumed that by automatically expanding the user's query with terms that are related to it, performance will increase. Clearly, the success of such techniques is closely tied to the quality of the thesaurus used. Various experimental results have shown that using general–purpose hand–crafted thesauri, such as WordNet, is usually not successful, even when manually disambiguating word senses

(e.g., [Smeaton and Berrut (1996)]). In general, such thesauri tend to be too shallow and broad to be useful. On the other hand, in domain–specific contexts, using a thesaurus that closely corresponds to the language use in the document collection of interest can significantly improve retrieval performance, but such thesauri are expensive to build and only available for a limited number of domains. As an alternative, it has been proposed to automatically build thesauri using the targeted document collection. The general idea is that term co–occurrence is a reasonable indication of semantic relatedness: terms which often occur in the same document (e.g., "service" and "tie–break") are likely related to the same concept (e.g., tennis). The main advantages of this method are that thesauri can be constructed without any cost, and, moreover, that the thesauri obtained are guaranteed to correspond closely to the document collection (e.g., contain the most significant terms from the collection). In [Qiu and Frei (1993)], an improvement of 20% was witnessed using such automatically constructed thesauri on a relatively small document collection; for larger collections, smaller improvements are usually found. However, the use of thesauri for query expansion is in practice heavily outperformed by a much simpler technique called relevance feedback [Billerbeck and Zobel (2004)]. This technique attempts to expand queries without the need for a thesaurus at all. Specifically, using the original query, a number of relevant documents are selected (e.g., the first 10 ranked documents). To expand the initial query, terms from these top–ranked documents are selected according to some criterion (i.e., terms occurring often in the top–ranked documents and relatively seldom in the document collection as a whole).

Finally, when moving to web search, a number of fundamental extensions to the general document retrieval model are needed. While using the web gives rise to a number of interesting opportunities, e.g., due to the existence of hyperlinks and standards such as HTML, at the same time it brings about new difficulties. For example, there is no quality control on the web: everybody can essentially publish anything. As a consequence, the relevance of a web document is not exclusively determined by its topic anymore, as in standard document retrieval models, but also by its quality or reliability. Therefore, search engines on the web combine scores for the topical relevance of a web page with scores estimating its quality. The best–known algorithm for estimating the quality of a web page is the PageRank algorithm [Brin and Page (1998); Page *et al.* (1998)], which uses hyperlinks to this end. The underlying assumption is that the more high quality web pages refer to a given page p, the more likely p is of high quality as