Information Extraction: Algorithms and Prospects
in a Retrieval Context

THE INFORMATION RETRIEVAL SERIES

Series Editor:
W. Bruce Croft
University of Massachusetts, Amherst

Also in the Series:

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** *by Gerald Kowalski;* ISBN: 0-7923-9926-9

CROSS-LANGUAGE INFORMATION RETRIEVAL, *edited by Gregory Grefenstette;* ISBN: 0-7923-8122-X

TEXT RETRIEVAL AND FILTERING: ***Analytic Models of Performance***, *by Robert M. Losee;* ISBN: 0-7923-8177-7

INFORMATION RETRIEVAL: UNCERTAINTY AND LOGICS: ***Advanced Models for the Representation and Retrieval of Information*****,** *by Fabio Crestani, Mounia Lalmas, and Cornelis Joost van Rijsbergen;* ISBN: 0-7923-8302-8

DOCUMENT COMPUTING: ***Technologies for Managing Electronic Document Collections***, *by Ross Wilkinson, Timothy Arnold-Moore, Michael Fuller, Ron Sacks-Davis, James Thom, and Justin Zobel;* ISBN: 0-7923-8357-5

AUTOMATIC INDEXING AND ABSTRACTING OF DOCUMENT TEXTS, *by Marie-Francine Moens*; ISBN 0-7923-7793-1

ADVANCES IN INFORMATIONAL RETRIEVAL: ***Recent Research from the Center for Intelligent Information Retrieval***, *by W. Bruce Croft*; ISBN 0-7923-7812-1

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** **Second Edition**, *by Gerald J. Kowalski and Mark T. Maybury;* ISBN: 0-7923-7924-1

PERSPECTIVES ON CONTENT-BASED MULTIMEDIA SYSTEMS, *by Jian Kang Wu; Mohan S. Kankanhalli;Joo-Hwee Lim;Dezhong Hong;* ISBN: 0-7923-7944-6

MINING THE WORLD WIDE WEB: ***An Information Search Approach***, *by George Chang, Marcus J. Healey, James A. M. McHugh, Jason T. L. Wang;* ISBN: 0-7923-7349-9

INTEGRATED REGION-BASED IMAGE RETRIEVAL, *by James Z. Wang;* ISBN: 0-7923-7350-2

TOPIC DETECTION AND TRACKING: Event-based Information Organization, *edited by James Allan;* ISBN: 0-7923-7664-1

LANGUAGE MODELING FOR INFORMATION RETRIEVAL, *edited by W. Bruce Croft; John Lafferty;* ISBN: 1-4020-1216-0

MACHINE LEARNING AND STATISTICAL MODELING APPROACHES TO IMAGE RETRIEVAL, *by Yixin Chen, Jia Li and James Z. Wang;* ISBN: 1-4020-8034-4

INFORMATION RETRIEVAL: ***Algorithms and Heuristics,*** *by David A. Grossman and Ophir Frieder,* 2nd ed.; ISBN: 1-4020-3003-7; PB: ISBN: 1-4020-3004-5

CHARTING A NEW COURSE: Natural Language Processing and Information Retrieval, *edited by John I. Tait;* ISBN: 1-4020-3343-5

INTELLIGENT DOCUMENT RETRIEVAL: Exploiting Markup Structure, *by Udo Kruschwitz;* ISBN: 1-4020-3767-8

THE TURN: Integration of Information Seeking and Retrieval in Context, *by Peter Ingwersen, Kalervo Järvelin;* ISBN: 1-4020-3850-X

NEW DIRECTIONS IN COGNITIVE INFORMATION RETRIEVAL, *edited by Amanda Spink, Charles Cole;* ISBN: 1-4020-4013-X

COMPUTING ATTITUDE AND AFFECT IN TEXT: Theory and Applications, *edited by James G Shanahan, Yan Qu, Janyce Wiebe;* ISBN: 1-4020-4026-1

Information Extraction: Algorithms and Prospects in a Retrieval Context

By

Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4987-0 (HB)
ISBN-13 978-1-4020-4987-3 (HB)
ISBN-10 1-4020-4993-5 (e-book)
ISBN-13 978-1-4020-4993-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

To those who search for meaning amidst ambiguous appearances...

Contents

Preface

Information extraction (IE) is usually defined as the process of selectively structuring and combining data that are explicitly stated or implied in one or more natural language documents. This process involves a semantic classification of certain pieces of information and is considered as a light form of text understanding. IE has a history going back at least three decades and different approaches have been developed. Currently, there is a considerable interest in using these technologies for information retrieval, since there is an increasing need to localize precise information in documents, for instance, as the answer to a question, rather than retrieving the entire document or a list of documents. Advanced retrieval models such as language modeling answer that need by building a probabilistic model of the content of a document. Question answering systems are trying to take the next step by inferring answers to a natural language question from a document collection. In these and other information retrieval models a semantic classification of entities, relations between entities, and of semantically relevant portions of texts (phrases, sentences, maybe passages) is very valuable to advance the state of the art of text searching. When talking about a semantic Web, semantic classification becomes of primordial importance, but also in other tasks that involve information selection and filtering, such as text summarization and information synthesis from different documents, IE is an indispensable preprocessing step.

The book gives an overview and explanation of the most successful and efficient algorithms for information extraction, and how they could be integrated in an information retrieval system. Special focus is on approaches that are fairly generic, i.e., that can be applied for processing heterogeneous document collections rather than a specific domain or text type and that are as much language independent as possible. The book contains a wealth of information on past and current milestones in information extraction, on necessary knowledge and resources involved in the extraction processes, and on the final aims of an extraction system. Additionally, a very important focus is on current statistical and machine learning techniques for information detection and classification. In an information retrieval context, these techniques can be used to learn and fine tune traditional knowledge engineered rules and patterns.

The book has grown from the results of a project on *Generic Technology for Information Extraction from Texts* (researched at the Katholieke Universiteit Leuven, Belgium) from 2000-2004 and sponsored by the Institute for the Promotion of Innovation by Science and Technology in Flanders) and from a graduate course on *Text Based Information Retrieval* taught at the same university to students in Artificial Intelligence, Informatics, and Electrical Engineering. This book is meant to give a comprehensive overview of the field of information extraction, especially as it is used in an information retrieval context. It is aimed at researchers in information extraction or related disciplines, but the many illustrations and real world examples make it also suitable as a handbook for students.

Acknowledgements

First, I would like to thank Rik De Busser, who is currently a Ph.D. student in Linguistics at La Trobe University in Melbourne, Australia, and who helped with the redaction of the first three chapters of this book. Secondly, I thank Prof. Jos Dumortier, the director of the *Interdisciplinary Centre for Law and Information Technology* at the K.U.Leuven for the opportunities given to our research group *Legal Informatics and Information Retrieval*. Many thanks go the staff of this group and especially to Roxana Angheluta, Jan De Beer, Koen Deschacht and Wim De Smet for participating in weekly project discussions. I am very grateful to Prof. Danny De Schreye, Head of the Informatics Department in the Faculty of Engineering for the many encouragements to pursue research in the domain of artificial intelligence. I sincerely thank Prof. Paul Van Orshoven, dean of our faculty, Prof. Yves Willems, former dean of the Faculty of Engineering, and Prof. Marc Vervenne, Rector of the K.U.Leuven, who gave me a marvelous chance to continue and perpetuate my research and teaching in the domain of information retrieval. Information extraction from written texts by a machine is a first step towards their automatic understanding. The task compares to decoding the symbols of an old language and gradually learning the meaning of the inscriptions. I am very grateful to the late Prof. Jan Quaegebeur (K.U.Leuven) and Prof. John Callender (University of California Los Angeles, USA). A long time ago they arouse in me the profound interest in content extraction from texts. I surely must thank Dr. Donna Harman (NIST, USA), Prof. Ed Hovy (University of Southern California, USA) and Prof. Karen Sparck Jones (University of Cambridge, UK) for creating influential and valuable ideas in the fields of information retrieval and text analysis. The final thank you goes to my family for their patience on Sunday afternoons.

1 Information Extraction and Information Technology

With Rik De Busser

1.1 Defining Information Extraction

A company wants to track the general sentiments about its newly released product in Web blogs. Another company wants to use the news feeds it bought from a press agency to construct a detailed overview of all technological trends in the development of semiconductor technologies. The company also wants a timeline of all business transactions involved in this development. A space agency allows astronauts to query large amounts of technical documentation by means of natural language speech. A government is gathering data on a natural disaster and wants to urgently inform emergency services with a summary of the latest data available. An intelligence agency is investigating general trends in terrorist activities all over the world. They have a database of millions of news feeds, minutes and e-mails and want to use these to get a detailed overview of all terrorist events in a particular geographical region in the last five years. A legal scholar is interested in studying the decisions of judges in divorce settlements and the underlying criteria. He or she has thousands of court decisions at his disposal. A biomedical research group is investigating a new treatment and wants to know all possible ways in which a specific group of proteins can interact with other proteins and what the exact results of these interactions are. There are tens of thousands of articles, conference papers and technical reports to study.

The above examples have a number of elements in common: (1) The requests for information; (2) The answer to this request is usually present in unstructured data sources such as text and images; (3) But, it is impossible for humans to process all data because there is simply too much of it; And

(4) computers are not able to directly query for the target information because it is not stored in a structured format such as a database but in unstructured sources. *Information extraction* (*IE*) is the subdiscipline of artificial intelligence that tries to solve this kind of problems.

Traditionally, information extraction is associated with template based extraction of event information from natural language text, which was a popular task of the *Message Understanding Conferences* in the late eighties and nineties (Sundheim, 1992). MUC information extraction tasks started from a predefined set of templates, each containing specific information slots that encode event types relevant to a very specific subject domain – for instance, terrorism in Latin America – and used relatively straightforward pattern matching techniques to fill out these templates with specific instances of these events from a corpus of texts. Patterns in the form of a grammar or rules (e.g., in the form of regular expressions) were mapped on the text in order to identify the information.

MUC was the first large scale effort to boost research into automatic information extraction and it would define the research field for the decades to come. Even at the time of writing, information extraction is often associated with template based pattern matching techniques. Unsurprisingly, the MUC legacy still resounds very strongly in Riloff and Lorenzen's definition of information extraction:

> *IE systems extract domain-specific information from natural language text. The domain and types of information to be extracted must be defined in advance. IE systems often focus on object identification, such as references to people, places, companies, and physical objects.* […] *Domain-specific extraction patterns (or something similar) are used to identify relevant information.*
>
> (Riloff and Lorenzen, 1999, p. 169)

This definition represents a traditional view on what information extraction is and it more or less captures what this discipline is about: The extraction of information that is semantically defined from a text, using a set of extraction rules that are tailored to a very specific domain. The main points expressed by this definition are that an information extraction system identifies information in text, i.e., in an unstructured information source, and the information that adheres to predefined semantics (e.g., people, places etc.). However, we will see in the rest of the book that at present the scope of Riloff and Lorenzen's definition has become too limited. Information

extraction is not necessarily domain specific. In practice, the domain of the information to be extracted is often determined in advance, but this has more to do with technological limitations of the present state of the art than with the long-term goals of the research discipline. An ideal information extraction system should be *domain independent* or at least portable to any domain with a minimum amount of engineering effort. Moreover, Riloff and Lorenzen do not specify further the types of information. Although many different types of semantics can be defined, the semantics – whether they are defined in a specific or a general subject domain – ideally should be as much as possible *universally accepted* and *bear on the ontological nature and relationships of being*.

Another consequence of the stress on pattern matching approaches that were developed during the MUC competitions is that eventually any technique in which pattern matching is used to organize data in some structured format can be considered to be information extraction. For instance, the early nineties saw a sudden surge in popularity of research into approaches that try to extract the content of websites (e.g., shopbots that extract and compare prices), usually in order to convert them into a more convenient, uniform structural format. Some of these approaches analyze the natural language content of full text websites, but many only use pattern matching techniques that exploit the structural properties of markup languages to harvest data from automatically generated web pages. While many researchers conveniently gathered these approaches under the common denominator *web based information extraction* (see for instance Eikvil, 1999), we will assume that information extraction presupposes at least some degree of semantic content analysis. In addition, information extraction is also very much involved in finding the relationships that exist between the extracted information, based on evidence in text (e.g., John **kisses** Claudia).

Cowie and Lehnert try to mend the previous inaccuracies. They see information extraction as a process that involves the extraction of fragments of information from natural language texts and the linking of these fragments into a coherent framework. In their view, information extraction

> *[…] isolates relevant text fragments, extracts relevant information from the fragments, and then pieces together the targeted information in a coherent framework. […] The goal of information extraction research is to build systems that find and link relevant information while ignoring extraneous and irrelevant information.*
>
> (Cowie and Lehnert, 1996, p. 81)

Cowie and Lehnert's interpretation of information extraction is close to what we need to solve the problems at the beginning of this chapter. There is still one thing missing in their definition. Although in this book we concentrate on information extraction from text, text is not the only source of unstructured information. Among these sources, it is probably the source where one has made the largest advancements in automatic understanding. But, other sources (e.g., image, video) exhibit a similar need for *semantically labeling unstructured information*, and advances in their automatic understanding are expected to occur in the near future. Any framework in which information extraction functions should not exclude this given.

The interpretations above are only a few representative definitions, and in the literature one finds additional variant definitions. To this multitude, we will add our own working definition, trying to incorporate the kernel task and function of information extraction and to avoid both Riloff and Lorenzen's and Cowie and Lehnert's limitations:

DEFINITION

Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks.

This definition is concise and covers exactly in what sense we will use the term *information extraction* throughout this book, but it is still fairly abstract. In the next sections, we will clarify what its constituent parts exactly mean.

1.2 Explaining Information Extraction

1.2.1 Unstructured Data

Information extraction is used to get some information out of *unstructured data*. Written and spoken text, pictures, video and audio are all forms of unstructured data. *Unstructured* does not imply that the data is structurally incoherent (in that case it would simply be nonsense), but rather that its information is encoded in such a way that makes it difficult for computers to immediately interpret it. It would actually be more accurate to use the

terms *computationally opaque* vs. *computationally transparent* data. Information extraction is the process that adds meaning to unstructured, raw data, whether that is text, images, video or audio. Consequently, the data become structured or semi-structured and can be more easily processed by the computer (e.g., in information retrieval, data mining or summarization).

In this book, the unstructured data sources we are mainly concerned with are *written natural language texts*. The texts can be of different type or genre (e.g., news articles, scientific treaties, police reports). Unless stated otherwise, we will assume that these texts are rather well formed, i.e., that they are largely coherent and error free. This is far from always the case. Written data, especially in an electronic context, is notorious for being incoherent and full of grammatical and spelling errors that are drafted with (e.g., spam messages) or without purpose (e.g., instant messages, postings on informal news groups). Errors might also occur in the output of an automatic speech recognition system. The algorithms described in this book can all be applied to these deviant types of textual data, provided that the system is specifically trained to deal with the characteristics of the relevant text type or that the probabilities of the translation into elements of well formed text are taken into consideration.

Limiting this book to information extraction from the text medium is no restriction of its value. The technologies described in this book contribute to an advanced understanding of textual sources, which, for instance can be used for aligning text and images when training systems that understand images. In addition, because technologies for the understanding of media other than text will be further developed in the near future, it seems valuable to compile information extraction technologies for text as they serve as a source of ideas for content recognition in other media or in combined media (e.g., images and text, or video).

1.2.2 Extraction of Semantic Information

Information extraction identifies information in texts by taking advantage of their linguistic organization. Any text in any language consists of a *complex layering* of recurring patterns that form a coherent, meaningful whole. This is a consequence of the *principle of compositionality* (Szabó, 2004), a general notion from linguistic philosophy that underlies many modern approaches to language and that states that the meaning of any complex linguistic expression is a function of the meanings of its constituent parts. An English sentence typically contains a number of constituent

parts (e.g., a subject, a verb, maybe one or more objects). Their individual meanings, ordering and realization (for instance, the use of a specific verb tense) allow us to determine what the sentence means. If a text would be completely irregular, it would simply be impossible for humans to make any sense of it.

It is yet not entirely clear how these linguistic layers exactly interact, but many linguistic theories and natural language processing assume the existence of a *realizational chain*. This theoretical notion has its roots in the grammar that was written by the Indian grammarian *Panini* in the 6th - 5th century B.C. (see Kiparsky, 2002). According to this notion, meaning in a language is realized in the linguistic surface structure through a number of distinct linguistic levels, each of which is the result of a projection of the properties of higher, more abstract levels. For instance, for Panini the meaning of a simple sentence starts as an idea in the mind of a writer. It then passes through the stage in which the event and all its participants are translated into a set of semantic concepts, each of which is in its turn translated in a set of grammatical and lexical concepts. These are in their turn translated into the character sequences that we see written down on a page of paper. Information extraction (and natural language processing, for that matter) assumes that this projection process is to a considerable extent bidirectional, i.e., that ideas are recoverable from their surface realizations by a series of inverse processes.

In other words, information extraction presupposes that although the semantic information in a text and its linguistic organization is not *immediately* computationally transparent, it can nevertheless be retrieved by taking into account surface regularities that reflect its computationally opaque internal organization. An information extraction system will use a set of extraction patterns, which are either manually constructed or automatically learned, to take information out of a text and put it in a more structured format. The exact techniques that are used to extract semantic information from a natural language text form the main topic of this book. Particular methodologies and algorithms will be discussed throughout its main chapters.

The use of the term *extraction* implies that the semantic target information is *explicitly present* in a text's linguistic organization, i.e., that it is readily available in the lexical elements (words and word groups), the grammatical constructions (phrases, sentences, temporal expressions, etc.) and the pragmatic ordering and rhetorical structure (paragraphs, chapters, etc.) of the source text. In this sense, information extraction is different from techniques that *infer* information from texts, for instance by building logical rules (logical inference) and by trying to distil world or domain

knowledge from the propositions in a text through deductive, inductive or abductive reasoning. We will refer to this latter kind of information as *knowledge*. Knowledge discovery is also possible by means of statistical *data mining* techniques that operate on the information extracted from the texts (also referred to as *text mining*). In all these operations information extraction is often an indispensable preprocessing step. For instance, information that is extracted from police reports could be used as the input for a data mining algorithm for profiling or for detecting general crime trends, or as the input of a case based reasoning algorithm that predicts the location of the next strike of a serial killer based on similar case patterns.

1.2.3 Extraction of Specific Information

Information extraction is traditionally applied in situations where it *is known in advance* which kind of semantic information is to be extracted from a text. For instance, it might be necessary to identify what kind of events are expressed in a certain text and at what moment these events take place. Since in a specific language, events and temporal expressions can only be expressed in a limited number of ways, it is possible to design a method to identify specific events and corresponding temporal location in a text. Depending on the information need, different models can be constructed to distinguish different kinds of classes at different levels of semantic granularity. In some applications, for example, it will suffice to indicate that a part of a sentence is a temporal expression, while in others it might be necessary to distinguish between different temporal classes, for instance between expressions indicating past, present and future.

Information extraction does not present the user with entire documents, but it extracts *textual units* or elements from the documents, typically simple or multi-term basic phrases (Appelt and Israel, 1999), which we also call *text regions*. As such, information extraction is different from *extractive summarization*, which usually retrieves entire sentences from texts that serve as its summary. Information extraction, however, can be a useful first step in *extractive headline summarization*, in which the summary sentence is further reduced to a string of relevant phrases similar to a newspaper headline.

Specificity implies that not only the semantic nature of the target information is predefined in an information extraction system, but also the *unit* and *scope* of the elements to be extracted. Typical *extraction units* for an extraction system are word compounds and basic noun phrases, but in some applications it might be opportune to extract other linguistic units,

such as verb phrases, temporal markers, clauses, strings of related meanings that persist throughout different sentences, larger rhetorical structures, etc. Whereas the unit of extraction has to do with the granularity of individual information chunks that are lifted out of the source text, the *scope* of extraction refers to the granularity of the extraction space for each individual information request. Information can be extracted from one clause or from multiple clauses or sentences spanning one or more texts before it is outputted by the system. Consider the example that an information question wants to retrieve event information about assassinations, it might be that the name of the person assassinated and the time and place of the event is named in a first sentence of a news article, but that the name of the assassin and his method are mentioned in some sentences further in the discourse.

During the Message Understanding Conferences (MUC), there gradually arose a set of typical information extraction tasks (see Grishman and Sundheim, 1996; Cunningham, 1997). A most popular task probably is *named entity recognition,* i.e., recognizing person names, organizations, locations, date, time, money and percents. These names are often expanded to protein names, product brands, etc. Other tasks are *event extraction*, i.e., recognizing events, their participants and settings, and *scenario extraction*, i.e., linking of individual events in a story line. *Coreference resolution*, i.e., determining whether two expressions in natural language refer to the same entity, person, time, place, and event in the world, also receives quite a lot of attention. These task definitions have been extremely influential in concurrent information extraction research and we will see that although they are getting too narrow to cover everything that is presently expected from information extraction, they still define its main targets. Currently, we see a lot of interest for the task of *entity relation recognition*. A number of *domain specific extractions* are also popular, e.g., extraction of the date of availability of a product from a Web page, extraction of scientific data from publications, and extraction of the symptoms and treatments of a disease from patient reports. The interest in the above extraction tasks is also demonstrated in the current *Automatic Content Extraction* (ACE) project.

1.2.4 Classification and Structuring

Typical for information extraction is that information is not just extracted from a text but afterwards also *semantically classified* in order to ensure its future use in information systems. By doing this, the information from unstructured text sources also becomes structured (i.e., computationally

transparent and semantically well defined). In the extreme case, the information that is verbatim extracted from the texts is discarded for further processing, but this is not what is usually intended.

Any classification process requires a semantic classification scheme, i.e., a set of semantic classes that are organized in some relevant way (for instance in a hierarchy) and that are used to categorize the extracted chunks of information into a number of meaningful groups. A very large variety of semantic classification schemes are conceivable, ranging from a small set of abstract semantic classes to a very elaborate and specific classification.

Based on the general information focus of a system, we can make a main distinction between *closed domain* and *open domain* (or *domain independent*) information extraction systems. Traditionally, information extraction systems were closed domain systems, which means that they were designed to function in a rather specialized, well delineated knowledge domain (and that they will therefore use very specific classification rules). For instance, most MUC systems covered very limited subjects such as military encounters, Latin-American terrorism or international joint ventures (Grishman and Sundheim, 1996). Domain independent information extraction systems, on the other hand, are capable of handling texts belonging to heterogeneous text types and subject domains, and usually use very generic classification schemes, which might be refined, if the information processing task demands a more specific identification of semantic information. The technology described in this book applies to both closed and open domain information extraction.

We mentioned before that information extraction essentially converts unstructured information from natural language texts into structured information. This implies that there has to be a predefined structure, a representation, in which the extracted information can be cast. Although the extracted information can solely be labeled for consequent processing by the information system, in the past many template based extraction systems have been developed. Template representations were typically used to describe single events (and later also complex scenarios) and consist of a set of attribute-value pairs (so-called *slots*), each of which represents a relevant aspect of the event (e.g., the action or state, the persons participating, time, place). An information extraction task traditionally tries to take information from a source text and maps it to an empty value of the defined template.

In order to know which piece of information is supposed to end up in which template slot, an information extraction application uses a set of extraction rules. These rules state which formal or linguistic properties a particular chunk of information must possess to belong to a particular

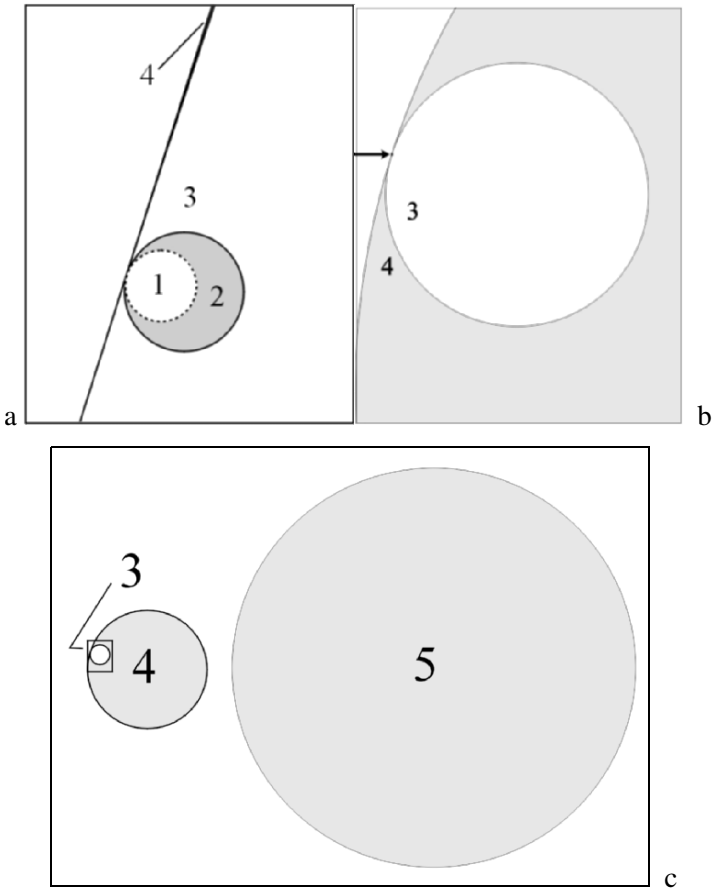
semantic class. Especially in earlier systems these rules were usually handcrafted (e.g., the FASTUS system developed by Appelt et al., 1993). Currently, machine learning is playing a central role in the information extraction paradigm. In most cases, *supervised learning* is used, in which a learning algorithm uses a training corpus with manually labeled examples to induce extraction rules, as they are applicable to a particular language and text type (e.g., the CRYSTAL system developed by Soderland et al. 1995). In some cases, it is also possible to apply *unsupervised learning*, for which no training corpus is necessary. For instance, unsupervised learning systems have been implemented for noun phrase coreferent resolution (e.g., Cardie and Wagstaff, 1999). Today, we see a large interest in weakly supervised learning approaches that limit the number of examples to be manually labeled (Shen et al., 2004). The application of these learning techniques has been one of the main enabling factors for information extraction to move from very domain specific to more domain independent analyses. In addition, the machine learning techniques more easily allow modeling a probabilistic class assignment instead of a purely deterministic one.

1.3 Information Extraction and Information Retrieval

1.3.1 Information Overload

Our modern world is flooded with information. Nobody exactly knows how much information there is – or how one could uniformly measure information flows from heterogeneous sources – but Lyman and Hal (2003) estimate that the total amount of newly created information on physical media (print, film, optical and magnetic storage) amounted to some 5 exabytes in 2002, most of it is stored in digital format. This corresponds to 9,500 billion books or 500,000 times the entire Library of Congress (which is supposed to contain approximately 10 terabytes of information). According to their measures, the surface Web contains around 167 terabytes of information, and there are indications that the deep Web, i.e., information stored in databases that is accessible to human users through query interfaces but largely inaccessible to automatic indexing, is about 400 to 500 times larger. Lyman and Hal (2003) estimate it to be at least 66,800 terabytes of data. A large fraction of this information is unstructured in the form of text, images, video and audio. These gargantuan figures are already

outdates at the time of writing and are dwarfed by the amount of e-mail traffic that is generated, which according to Lyman and Hal (2003) amounts to more than 300,000 terabytes of unique information per year.



Legend	
1	Size of the Library of Congress
2	Size of the surface Web
3	Size of the surface + deep Web
4	Size of surface + deep Web + e-mail traffic
5	Size of text data on hard discs sold in 2003

Fig. 1.1. Graphical presentation of the size of the Web and global storage capacity on computer hard discs anno 2003.

According to these authors, during 2003 an estimated 15,892.24 exabytes of hard disc storage was sold worldwide. A similar study that confirms the information overload is made by O'Neill, Lavoie and Bennett (2003). If this trend continues we will have to express amounts of information in yottabytes (2^{80} number of bytes).

In order to give a rough impression about the amounts of data that are involved, Fig. 1.1 gives a graphical representation of the total amount of data present on the Web and on hard discs worldwide in 2003. Data ratios are reflected in the relative size differences between the diameters of the circles. Figure 1.1a shows the size of unique textual data on the surface web [2] in comparison with the textual data on the combined surface and deep web [3] and of the surface and deep Web plus all e-mail traffic [4]. The size of the Library of Congress [1] is given as a reference. Figure 1.1b is a 180-fold magnification in which Fig. 1.1a appears as the minute rectangle at the point of tangency of 3 and 4. Fig. 1.1c gives an impression of the complete size of the textual web at the left hand side and a comparison with all textual data on hard discs on the right hand side.

This immense information production prevents its users to efficiently select information and accurately use it in problem solving and decision making (Edmunds and Morris, 2000; Farhoomand and Drury, 2002). Even if we find ways for reducing the information generation, there still is a large demand for intelligent tools that assist humans in information selection and processing (Berghel, 1997).

1.3.2 Information Retrieval

Information retrieval (*IR*) is a solution to this kind of problem (Baeza-Yates and Ribeiro-Neto, 1999). It allows a user to retrieve a set of documents from large document collections, such as the Web or a corporate intranet, based on a keyword based query. Information retrieval is able to search efficiently through huge amounts of data because it builds indexes from the documents in advance in order to reduce the time complexity of each real-time search. The low level keyword matching techniques that are generally used in information retrieval systems make them error tolerant, domain independent and – above all – very fast (Lewis and Sparck Jones, 1996). The success of information retrieval systems in general, and the Web search engines in particular is largely due to the flexibility of these systems with regard to the queries that users pose. Users have all kinds of information needs that are very difficult to determine a priori. Because users do not always pose their queries with the words that occur in relevant documents, query expansion with synonym and related terms is very

popular, primarily enhancing the recall of the results of the search. Information retrieval is very successful in what it is aimed to do, namely providing a rough and quick approach to find relevant documents.

A downside is that such a robust and flexible approach sometimes results in a low precision of an information search and in a huge number of possibly relevant documents when a large document base is searched, which are impossible to consult by the user of the information system (Blair, 2002).

1.3.3 Searching for the Needle

Because of the information overload, the classical information retrieval paradigm is no longer preferable. This *paradigm* has found its roots in the example of the *traditional library*. One is helped in finding potentially relevant books and documents, but the books and documents are still consulted by the humans. When the library is becoming very large and the pile of potentially relevant books is immensely high, humans want more advanced information technology to assist them in their information search. We think that information extraction technology plays an important role in such a development.

Currently, an information retrieval system returns a list of relevant documents, where each individual document has to be fetched and skimmed through in order to assess its real relevance. There is a need for tools that reduce the amount of text that has to be read to obtain the desired information. To address this need, the information retrieval community is currently exploring ways of pinpointing highly relevant information. This is one of the reasons *question answering systems* are being researched. The user of a question answering retrieval system expresses his or her information need as a natural language question and the system extracts the answer to the information question from the texts of the documents (Maybury, 2003).

Information extraction is one of the core technologies to help facilitate highly focused retrieval. Indeed, recognizing entities and semantically meaningful relations between those entities is a key to provide focused information access.

With the current interest in expressing queries as natural language texts, the need for semantic classification of entities and their relations in the texts of document and query becomes of primordial importance. Information extraction technology realizes that – simply saying – not only the words of the query, but also the semantic classifications of entities and

their relations must match the information found in the documents (Moens, 2002).

Especially in information gathering settings where the economical costs of searching is high or in time critical applications such as military or corporate intelligence gathering, a user often needs very specific information very quickly. For instance, an organization might need a list of all companies that have offices in the Middle East and conducted business transactions or pre-contract negotiations in the Philippines or Indonesia in the last five months. He or she knows that many of this information is available in news feeds that were gathered over the last half year, but it is impossible to go through tens of thousands of news snippets to puzzle all relevant data together. In addition, we cannot neglect the need for flexible querying. There will always be a large variety of dynamically changing information needs.

Information retrieval techniques typically use general models for processing large volumes of text. The *indices* are stored in data structures that are especially designed to be efficiently searched at the time of querying. An ideal information retrieval system answers all kinds of possible information questions in a very precise way by extracting the right information from a (possibly large) collection of documents.

Information extraction helps building such information systems. The extracted information is useful to construct sensitive indices more closely linked to the actual meaning of a particular text (Cowie and Lehnert, 1996). This is often only restricted to the recognition and classification of entities that are referenced in different places of the text and recognition of relations between them. Besides an index of words that occur in the documents, certain words or other information units are tagged with additional semantic information. This meta-information allows more precisely answering information questions without losing the advantages of flexible querying. Information extraction technology allows for a much richer indexing representation of both query and document or information found in the document, which can improve retrieval performance of both open and closed domain texts. Especially linguistically motivated categories of semantics become important (e.g., expressions of time, location, coreference, abstract processes and their participants, ...). As we will show in this book, the identified and classified information – even if very generic semantic classifications are made – is useful in information retrieval and selection, allowing for answers of information needs that can be more precisely inferred from information contained in documents. Information extraction can be regarded as