Information Retrieval Techniques

Faculty of Philosophy, Department of Information Sciences

Jadranka Lasic-Lazic, Sanja Seljan, Hrvoje Stancic

Abstract

There is currently huge amount of data on the Web and almost no classification information. The key problem is how to embed knowledge into information mining algorithms. The authors analyse techniques of information retrieval and give their strong and weak points.

Although most Web documents are text oriented, there are plenty of them that contain multimedia elements, which are not easily accessible through common search methods. Web information is dynamic, semi-structured, and interwound with hyperlinks. Several advanced methods for Web information mining are analyzed: 1) syntax analysis, 2) metadata-based searching using RDF, 3) knowledge annotation by use of conceptual graphs (CGs), 4) KPS: Keyword, Pattern, Sample search techniques, and 5) techniques of obtaining descriptions by fuzzification and back-propagation.

The problem of choosing proper keywordsis also stressed out. The authors suggest the usage of already accepted standards for classification hierarchy, such as Dewey Decimal Classification (DDC).

1 INTRODUCTION

As greater volumes of documents including multimedia elements became available on the Internet, users need more sophisticated tools to locate the information that is relevant to them. Therefore, a number of language technologies are being deployed in variety of information management applications (ref. 3): multilingual search engines, Machine Translation (MT), video access system, content-based language technologies for information systems, document summarization, robust text processing for document content abstraction, E-commerce, document/database relevancy visualization, Web browser instrumentation, information extraction, human-computer dialog systems and multimodal interfaces to content visualization.

The key problem of this article is to analyze problems of embedding knowledge into information mining algorithms and finding suitable techniques for information retrieval. There are large collection of multimedia documents and lexical databases built in a shared mark-up language. However, the creation of standards in encoding the knowledge is essential for ensuring integration and use. The European Commission and the National Science Foundation support development of new language technologies for Web applications through a variety of activities (http://www.linglink.lu/hlt/call-for-proposals/). European Commission also reports that non-English Web material is growing faster than the English material (in 1997, 75% of commercial Web sites in Europe are other than English), and Forrester Research Inc. reports that 82% of Web sites for Europe's largest firms offer content in more than one language.

It is obvious that rapid growth of economic and social development require urgency to convey and access information in different languages. There is growing need for Web-based technologies which facilitate information access and dissemination by bridging linguistic, social and geographic barriers require integration, development and standardization of information technologies emerging from the areas of computing, informatics, telecommunications, language engineering and economy.

2 INFORMATION RETRIEVAL TECHNIQUES

Web pages mostly contain semi-structured and dynamic information, interwound with links and not easily accessible. Searching through WWW is significantly different from searching data in databases, which are static and centralized. It has been developed a number of query languages, based on semi-structured data models and mostly represented as labeled graphs.

The key problem is how to embed knowledge into information mining algorithms. Although most of Web documents are text-oriented, considerable amount of information is not easily accessible through common search methods, so documents can't be retrieved without accessing each one individually.

Several advanced methods for Web information mining are analyzed: 1) Syntax analysis, 2) Metadata-based search using RDF (Resource Description Framework), 3) Knowledge annotation by use of CGs (Conceptual Graphs), 4) KPS: Keyword, Pattern, Sample search techniques, 5) Techniques of obtaining descriptions by Fuzzification and Back-propagation. The problem of choosing proper words and indexation is also stressed out.

2.1 Syntax analysis

Full-text search is probably the best known method, performing string-matching (e.g. using regular expressions) and structure-matching searches (e.g. tags, link names and link paths) in documents. Internet search engines such as Altavista, Infoseek, Excite etc., i.e. Web robots construct index of key words found in documents trying to capture in that way the content of documents. Queries can be refined by using logic operators (AND, OR, NOT). The general advantage is to be fast due to automated indexing, where no human intervention is needed. Disadvantage is that answers are often irrelevant, incomplete or the number of results may be very large and, therefore, not usable.

One way to improve information retrieval is to use knowledge representation language (KRL) in order to index Web documents. One of them is RDF - Resource Description Framework, built over XML which is more machine-readable than human readable format. Another way to ease representation is to use set of intuitive and combinable commands equivalent to first-order logic, such as Conceptual Graphs (CGs) for indexing any Web information.

2.2 Metadata-based search (RDF)

Metadata-based search is based on meta-descriptions given form documents. Attributes should describe properties and content of the document. There are two types of them (ref. 8): basic type similar to the system used in libraries with predefined set of attributes (e.g. author, title, ISBN) and intelligent type that can deduce information from semantic network.

A major problem is that there is no reliable method for finding out relevant document without accessing each one individually. Therefore, WWW Consortium has introduced RDF in order to produce standard language for machine-readable description of resources on the Web.

Automatic metadata generation, according to Dewey Decimal Classification, that classifies HTML documents, can be used to extract context sensitive metadata which is then represented using RDF (ref. 5). Automatic classifier is an object oriented system, written in Java, that retrieves HTML document form given URL, analyses the content and assigns appropriate DDC classification classmark. It compares terms found in the document with manually defined terms representing nodes of DDC hierarchy. The result is useful metadata such as document title, keywords, abstract and word count. Documents sharing the same subject matter will be clustered under the same classification mark.

RDF is seen as "Web of trust" where each document should be well described and universally understood. Various attempts have been made to include metadata into HTML documents. The problem is that such technique is not obligatory. Contrary to this idea, M. Marchiori proposes back-propagating meta information from pages with known metadata to those that are linked form (see 2.6).

2.3 Conceptual Graphs (CGs)

P. Martin and P. Eklund (ref. 7) argue in favor of general knowledge representation languages for indexing Web documents and believe that they have advantages over metadata languages based on Extensible Mark-up Language (XML). Information retrieval is better supported by languages designed to represent semantic content. Therefore, they suggest the use of concise and easy comprehensible *Conceptual Graphs* (CGs).

Conceptual graphs are formalized in an abstract syntax that is independent of any notation, but each of them can be defined in the graphical *display form* (DF), formally defined as *conceptual graph interchange form* (CGIF), and as readable *linear form* (LF). CGs can be translated in logically equivalent predicate calculus and Knowledge Interchange Format (KIF), (ref. 2).

They suggest to use set of intuitive, complementary and combinable language commands that allow indexing of Web pages, at the levels of precision they desire. It is possible for the user not to declare all the terms. This semi-formal model is at expense of knowledge precision and accessibility. The WebKB accepts CGs that include undeclared terms, but then the convention of using basic relations is important.

There are several ontologies proposed for knowledge representation. The ontology that proposes WebKB is more terminologically oriented to ease rapid and simple knowledge representation. It includes 400 concepts and relations (e.g. thematic, spatial, temporal relations etc.) and was created by merging other top-level ontologies used in knowledge acquisition, knowledge representation and in cooperation with hypertext tools. The Word Net ontology (120.000 words linked to 90.000 concepts) is included into this top-level ontology.

Lexical, structural and knowledge-based techniques are combined to retrieve or generate knowledge or Web documents. In a WebKB, lexical and structural query commands are proposed and may be combined with knowledge query commands. It is important that all commands can be embedded within documents. Although ontological distinctions seem obvious, the user may often make semantic errors when representing knowledge. Therefore, WebKB should perform some checking when the user performs classification. At present WebKB exploits only CG formalisms (ref. 7).

Opposed to XML, which is machine-readable rather than human readable, and to RDF which is XML-based metadata language characterized by poor readability, WebKB proposes the use of expressive but intuitive KB representation languages to index information in documents. To allow this, the knowledge must be enclosed within HTML tags <KR language = "CG">. Thus, there is no need to separate knowledge from documentation. Another facility of WebKB is that it ignores HTML tags (except definition list tags), so HTML or XML features can be combined with knowledge statements. Web KB also allows *to index an image* by a knowledge statement, or to isolate *any* textual/HTML data as a Document Element to indexed by knowledge statements.

P. Martin and P. Eklund enumerate three advantages of semantic network structure of CGs:

(a) restricted formulation of knowledge

(b) user is encouraged to formulate relations between concepts

(c) better visualization of relations between concepts.

Current trend is to allow users to annotate documents using metadata languages. In that way users can represent and query documents at the desired level. The notation of Conceptual Graphs is simple and leaves some terms undeclared. To support this approach, WebKB ontology is suggested to support building of Web-accessible storage of knowledge.

2.4 KPS: Keyword, Pattern, Sample search techniques

According to T. Guan and K.-F. Wong (ref. 4) the method of syntax analysis of HTML tags is adequate only for highly structured Web-pages, and the use of wrappers or user-defined declarative languages is time-consuming and not suitable for thousands of information resources. Therefore, they suggest a new algorithm, KPS standing for *Keyword, Pattern, Sample* for information mining.

Most of Web pages are text-oriented, but contain information embedded in the text, e.g. the biography relations (TIME, DEGREE, SCHOOL), that cannot be specified easily. Sometimes the data is not tagged and sometimes the employed tags are different even if the items are similar or the same.

T. Guan and K.-F. Wong propose KPS algorithm for extracting information from Web pages. The goal is to extract information from irregular pages automatically or with minimal human efforts, using *keywords, patterns and/or samples*. There are several assumptions for using this algorithm:

(a) Important information is always highlighted by keywords or meaningful structures

(b) Common patterns exist in many languages, e.g. *Mr.* or *Dr.* followed by a name

(c) Similar structures or patterns exist usually in the same organization and are often designed by the same person

Information extraction (IE) requires domain-specific knowledge, and is quite different from quering data in traditional databases, since it uses semi-structured features of Web pages. Here are basic ideas of KPS mining algorithm:

1) *Keyword-based mining* is used to extract value related to a keyword, which can be link to the next page, word in the title, an item of the list, field in the table or ordinary word in the text. Synonyms and hyponyms are also considered in the text, e.g. *E-mail* include also *e-mail, email, father* is a hyponym of *parent*, and *parent* is hyponym of *relative* etc. Dictionary is based on WordNet ontology. T. Guan and K.-F. Wong suggest to use keyword-based mining for *publications, research interests, E-mail* etc. since most of

them are highlighted as keywords, but not for searching the label as *Name* which doesn't usually stand before the proper name.

2) *Pattern-based mining* performs string matching based on patterns which are specified by users. A pattern consists of constant words or variables (started with /) enclosed in a pair of brackets. For example, the pattern [Dr. /*Name*] matches with a string starting with *Dr.* and followed by a noun phrase, the variable *Name*. Therefore, exact name will be assigned to the variable *Name*.

   Two or more patterns may be linked using Boolean operators [Dr. /*Name*] or [Ph.D. /*Name*]. It is possible to use signs * and – to denote any number of words or a word, respectively or e.g. search with complex pattern [Dr. /*Name* received /*Degree* from * in /*Year*]. T. Guan and K.-F. Wong suggest to use pattern-based mining for *E-mail, telephone, address, professors, doctors* etc. since they almost have the same pattern. The pattern [*@*.*] cannot detect all names in an institution since some persons do not have e-mail, or it is hidden under the link *contact* or the title is not included in the homepage.

3) *Sample-based mining algorithm* extracts information based on a sample, defined by a user. Supposing that several Web pages are written by the same author, the user who wants to find e-mails in an institution can locate manually and the system will help automatically. Guan and Wong give precise algorithm for retrieving *pattern and style similarities* (ref. 4). There are several suppositions for sample-based mining algorithm:

   (a) *Web pages consist of a list of fields* $w_i=(f_1, f_2, ..., f_i)$, where each field can be a word (conference, WWW8), number (4.56, 11), a date (2000, 15/06/87, 15-06-97, July 15), time (10:35, 8 am), price ($45, HK40), specific ASCII characters except letters 'a'...'z' and 'A'...'Z', numbers '0'...'9' and '+',' -' etc. and HTML tags <br>, </br>.

   (b) *An object* ***o*** *is a list of continuous fields* appearing in the body, where the first and the last elements cannot be HTML tags. The person's name may be represented as an object. *Sample* is a specific object indicated by the user.

   o = (Tracy, Hanon)

   There are two types of similarities between sample and potential objects: pattern similarity and style similarity.

T. Guan and K.-F. Wong suggest to use sample-based mining to find information containing almost the same pattern, e.g. personal Web pages often share some common structures. Most professor's homepages have *name, title, biography, teaching, research interests* and *publications*.

Although the suggested pattern cannot mine all desired information, the KPS algorithm is useful to extract text-oriented semi-structured Web pages. Extraction, e.g. *rates of hotel for single-room or double room, adults and children* require semantic knowledge and NLP techniques. Using of KPS algorithm cannot guarantee 100% success in mining the desired information, but initial experience shows it is practical.

2.5 Fuzzification

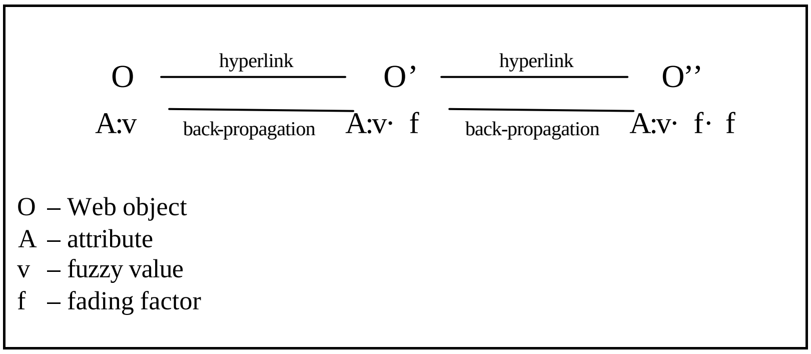
It is a common fact that information retrieval of the desired information from the Web can be a tiresome process. The main reason is the poor classification of the Web information. Of course, there are plenty of Web search engines which utilize special robots in search for new Web pages, and when a page is found it is put in the 'right' classification category depending on the classification method the Web search engine is using.

On the other hand, metadata classification can be added to Web objects. This means that the task of classification is partially transferred to those who create and maintain Web elements. As an advanced solution for Web classification M. Marchiori (ref. 6) proposes the fuzzification method. He says that existing Web metadata sets do have attributes assigned to objects, but they either have them or do not have them. Instead, he argues that attributes should be *fuzzified*, i.e. each attribute should be associated with a "*fuzzy measure* of its relevance for the Web object, namely a number ranging from 0 to 1". This means that if an attribute is assigned value 0 it is not pertinent to the correspondent Web object. If the value is 0.4 relevance of the attribute to the Web object is 40%. Since classification by itself is an approximation, better of worse, fuzzification method allows flexibility within a predefined classification system providing more detailed ranking and allowing the basic set of concepts to be relatively small.

2.6 Back-propagation

So far we have seen how to calculate relevance of a Web object to the search query providing the information that is contained within the Web object. Since Web is a dynamic media interwound with hyperlinks it is a common fact that one Web object points to some other Web object or even to more of them. This brings us to the problem of calculating relevance of an object that is pointed to by some other object, as described in the following model by M. Marchiori (ref. 6).

Suppose a certain Web object O has the associated metadatum A:v, indicating that the attribute A has fuzzy value v. If there is another Web object O' with an hyperlink to O, then we can "back-propagate" this metadata information from O to O'. The intuition is that the information contained in O (classified as A:v) is also reachable from O', since we can just activate the hyperlink. However, the situation is not like being already in O, since in order to reach the information we have to activate the hyperlink and then wait for O to be fetched. So, the relevance of O' with respect to the attribute A is not the same as O'(v), but is in a sense faded, since the information in O is only potentially reachable from O', but not directly contained therein. The solution to this problem is to fade the value v of the attribute multiplying it by a "fading factor" f (with $0<f<1$). So, in the above example O' could be classified as $A:v\cdot f$. The same reasonment is then applied recursively. So, if we have another Web object O'' with an hyperlink to O', we can back-propagate the obtained metadatum $A:v\cdot f$ exactly in the same way, obtaining that O'' has the corresponding metadatum $A:v\cdot f\cdot f$.



Experiments on a randomly chosen region of Web objects showed that the usage of back-propagation method can significantly improve effectiveness of the classification. They also showed that the critical mass of Web metadata classification usefulness is achieved when at least 16% of the Web use metadata classification, in contrast with 50% without incorporation of the back-propagation method. Furthermore, in order to achieve excellence level metadata need 53% of the Web to be classified, in contrast with 80% without described method.

Most of all, the method of back-propagation, which presuppose the fuzzification method, acts on top of any classification, and does not require any form of semantic analysis. Therefore, it is completely language independent which is very important when the number of non-English Web pages is constantly increasing.

3 INDEXING AND INFORMATION RETRIEVAL

Before the description of an indexing model is given, the internal organization of an information retrieval (IR) model should be explained. According to Y. Chiaramella (ref. 1) IR models consists of:

- a model of documents (Web objects),
- a model of queries,
- a matching function which compares queries to documents (Web objects), and
- a knowledge base.

There are several kinds of knowledge that IR model has to tackle with. *Content knowledge* consists of the domain concepts that describe the semantic content of basic Web objects, or atomic data. *Structural knowledge* is made of links between basic Web objects. The concept of *domain knowledge*, however, consists of the two mentioned types of knowledge. Some of those knowledge may be implicit and therefore the knowledge base is used to make it explicit. Within a knowledge base indexing process, as a form of information mining, identifies and extracts implicit knowledge from information.

In order to better understand indexing model Y. Chiaramella defines *index units* as a structural units that are indexed (i.e. that are assigned an explicit representation of their semantic content) and consequently units that are individually retrievable from queries that include content requirements. The semantic content of different units may not be independent. This is particularly obvious in the case of linked nodes. Therefore, in order to achieve the desired level of hierarchy organization, indexing should follow the logical structure of a document. Every document should have defined its minimum and maximum index unit, attribute value, internal references between nodes of the document, external references towards nodes of other documents, etc.

There are two types of attributes:

1) *Dynamic attributes* which propagate their value in the logical structure, and can be:

- Descending – propagate value from top to bottom of the hierarchical logical structure (e.g. publication date), and
- Ascending – propagate value from bottom to top of the hierarchical logical structure (e.g. author, and if there is more than one, than the logical whole is co-authored);

2) *Static attributes* which do not propagate their values, but correspond to properties that remain purely local to the structural object they are assigned to (e.g. title).

Indexing model, with its hierarchically structured units, showed as isomorphic to the hierarchy of abstraction levels in the Conceptual Graphs method of information mining.

4 CONCLUSION

Current information retrieval techniques cannot give precise results, because of not highly structured Web pages, which are dynamic, semi-structured and contain multimedia information. Current trend is to allow users to annotate documents using metadata languages.

The problem of proper indexing and subjective classification is very important, so universally known classification is recommended. All Web documents cannot be classified manually.

As an intermediate step, there is a trend to annotate documents, using metadata languages: XML-based or CGs. The notation of Conceptual Graphs is simple, intuitive, can annotate any web information (including picture) and leaves some terms undeclared. In that way the users can represent and query the documents at desired level. WebKB ontology is suggested to support building of the Web-accessible storage of knowledge.

Although it is very hard to automate resource description, automatic metadata generation seems to be essential requisite enabling description of any HTML page. Disadvantage is also that is more machine-readable than human-readable format and not precise enough.

Usage of KPS algorithm is probably more suitable for searching one site, than the whole Web. Although it cannot mine all desired information, is very useful for information extraction of textual Web pages.

On the other hand, methods of fuzzification and back-propagation aid existing classification and are relying only on the interconnectivity of the Web pages. They are applied on top of the classification and therefore are language independent.

Indexing process within a knowledge base is applied to the internal logical structure of a document in order to retrieve implicit knowledge from information.

Current information retrieval techniques cannot give precise answers about semantic content of documents, because of difficulties in automated extraction of knowledge. Therefore, more work should be done to apply semantic knowledge and natural language processing techniques.

Jadranka Lasic-Lazic, PhD. is professor at the Department of Information Sciences, Faculty of Philosophy, University of Zagreb. She is teaching Theory of Classification, Classification and Classification Systems, and Indexing and Retrieval systems. She is a member of International Reading Association and Croatian Librarianship Association.
E-mail: jlazic@mudrac.ffzg.hr

Sanja Seljan, MSc. is assistant at the Department of Information Sciences, Faculty of Philosophy, University of Zagreb at the project "Machine undestanding of the Croatian language". Her research interests are Natural Language Processing (NLP), Lexical- Functional Grammar (LFG), and Machine Translation (MT).
E-mail: sseljan@ffzg.hr

Hrvoje Stancic, B.A. is working at the Department of Information Sciences, Faculty of Philosophy, University of Zagreb. He is currently a M.A. student of Information Sciences. His research interests are Document and Knowledge Management, and Information Storage and Retrieval.
E-mail: hrvoje.stancic@zg.tel.hr

References

1) Chiaramella, Yves, *Browsing and Querying: two complementary approaches for Multimedia Information Retrieval*, CLIPS Laboratory, Grenoble, France, http://eldorado.uni-dortmund.de:8080/FB4/tagung/97/ HIM97/paper1, July 23, 2000

2) Conceptual graph Examples (http://www.bestweb.net/~sowa/cg/cgexampw.htm)

3) Frederking, R. et al., *Language on trial*

4) Guan, Tao, Wong, Kam-Fai, *KPS: a Web information mining algorithm*, University of Regina, Canada and The Chinese University of Hong Kong, China, http://decweb.ethz.ch/WWW8/data/2174/html/ index.htm, April 29, 2000

5) Jenkins,C. et al., *Automatic RDF metadata generation for resource discovery* http://decweb.ethz.ch/ WWW8/data/2138/html/index.html

6) Marchiori, Massimo, *The limits of Web metadata, and beyond*, MIT Laboratory for Computer Science, USA, http://decweb.ethz.ch/WWW7/1896/com1896.htm, April 29, 2000

7) Martin, P., Eklund, P., *Embedding knowledge in Web documents*, Griffith University, Australia, http://decweb.ethz.ch/WWW8/data/2145/html/bindex.htm, April 24, 2000

8) Stanoevska, K. et al. Efficient Information Retrieval: Tools for Knowledge Management. In: Reimer, U.: Practiac Aspects of Knowledge Management PAKM 98. Proceedings of the Second International Conference in Basel, Switzerland, October, 1998. http://www.knowledgemedia.org/netacademy/ publications.nsf/all_pk/1137