Information mining and information retrieval : methods and applications

J. Mothe, C. Chrisment

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 Route de Narbonne,
31062 Toulouse Cedex, France
Fax : 33 5 61 55 68 52 Tel : 33 5 61 55 63 22
e-mail : mothe@irit.fr , http://www.irit.fr/~Josiane.Mothe

Abstract

New generations of systems have to be designed in order to exploit the increasing masses of multimedia data available electronically. In addition to efficient retrieval engines that give direct access to document contents, the users need some tools that help them to discover information, derive new information from sub-sets of documents. This paper introduces how information mining can help information retrieval.

1 Introduction

The amount of information available electronically is so huge that new generation of systems has to be designed to exploit it. In order to provide efficient information retrieval (IR), new types of engines have to be designed so that the stored information is organised according to the semantics of the document contents and to the users' behaviour or information usage. Moreover, in addition to efficient retrieval engines that give direct access to the document content, the users need some tools that provide them with global views of the available information, help them to discover information in huge mass of documents and derive new information or knowledge from target sub-collections of documents.
Information mining (IM) contributes responses to these new types of information needs. Two complementary types of information can be considered: content and usage.
When considering **content mining**, the studied object is the message carried by the document contents. Texts for example can be considered with regard to its content or informative components rather than an unorganised bag of words or descriptors giving access to it. In that case, the goal of text mining is message understanding as it is done in MUC[1] (a single document can be considered) or more generally extraction of domain knowledge as terminology or ontologies (from a target set of documents). In turn, this extracted knowledge can be used in order to organise the document collection and to guide the user when browsing or searching the collection. Alternatively, documents can be considered as composed of typed pieces of information and target elements can be extracted (objects or characters in a picture or in a video, patterns in a string of music, organisations in texts, etc.). Among these pieces of information, meta-information plays a core role but any concept occurring in a document or a set of documents can play such a role as soon as it can be associated with a semantic or with a user's usage. In a step further, these elements can be used to extracted correlations and to derive new information

[1] Message Understanding Conference, www.cs.nyu.edu/cs/faculty/grishman/muc6.html

(trends, time dependence, etc). New standards (XML, DAML/OIL, MPEG7) are developed to help the automatic content mining.
In a complementary way, **usage mining** is based on analysing the users behaviour (logins, browsing, queries, domains of interest) and is used to extract users' preferences in term of content, interface, etc. (Chang, 2001). Customize systems can then be designed to fulfil users' preferences: providing document suggestions, navigation propositions or personalised interfaces.

2 Applications

A vast range of research domains and applications are involved in IM for IR. We focus on the ones we believe are the most important for the next 10 years and on which some research has already been carried out and provide some solutions even if there are still partial.

2.1 Knowledge representation - ontologies

Definitions in computer science insist that ontologies are generic and conceptual descriptions of the domain entities required to design a knowledge-based application (ontology, 2002). They provide a sound basis for communications between human and machine agents. The objective of "defining meaning" urges us to thoroughly distinguish between objects and beings, the symbols that stand for them (words or phrases) and the representations required by an agent that uses this symbol to evoke this being. Concepts correspond to these representations, whether mental or formal. The notion of ontology has evolved quickly: in spite of on-going interest for general ontologies with a universal scope such as WordNet (www.cogsci.princeton.edu) or Sensus (www.isi.edu), ontologies often now refer to domains that restrict their scope.
What are the contributions of ontologies to the IR process?

- Ontologies should enhance retrieving effectiveness and the user should express his needs more easily: an ontology should be precise enough to provide a unique definition for terms, to deal with synonymy and ambiguity, to look for some concepts that are not explicitly written in the user's queries or in the information to annotate,
- Ontologies should facilitate the IR from various heterogeneous knowledge sources,
- Ontologies should be a key component of the Semantic Web. "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." (Berners, 2001).

2.2 Automatic summarisation

A summary is a brief account giving the main points of a whole document. It has been used in IR mainly to provide a synthetic view of a retrieved document and help the user to decide on the relevance of a given document.
Text document summarization has largely been investigated and efficient solutions are provided to build summaries either written in natural language or consisting of pre-defined elements of information that give the main points of the document (e.g. person or organization names, dates).
Automatic summarisation of other type of media has been investigated recently and the solutions are still in its infancy: one key picture can be used to summarize a sequence of video. More advance information summarisation should be defined in order to help users to know what is inside a given document. A step further would be to summarize an entire collection taking into account what the user is looking for.

2.3 Information categorisation and clustering

Categorisation and clustering techniques aim at grouping together similar documents (Chakrabarti,03). These two types of techniques differ in how the groups are built. *Categorisation* refers to methods in which pre-defined categories are set and documents have to be associated with, generally, one category. Approaches in which a document can be associated with several categories are now defined too. The common principle consists to build a classifier by learning from a set of pre-categorized documents and defining the profile of each category. With regard to textual documents, the profiles are defined by a set of (weighted) terms. In the same way, images can be categorised according to their characteristics (colours, texture, shape, and land cover). *Classification* methods are based on inner similarities of classified objects.
Current approaches focus in grouping homogeneous collections of objects and a few investigate multimedia sources that should combine different classifier and however provide the user with a unified classification. Additionally, new ways of classifying information should be more user-oriented (taking into account the information usage, why the user search for the information, who is the user: his level of knowledge of the domain and which level of knowledge he needs).

2.4 Information monitoring and competitive intelligence

Competitive intelligence is a core issue in the new world of business. Companies need methods and tools to monitor the activities of its competitors, get information on the market, technologies, or government actions. These spying activities are necessary for them to define alliance strategies, innovation and customer oriented strategies.
In IR, citation and co-authoring analysis have been studied in the past as a way to monitor scientific activities. In the web context, hypertext references are mined to determine the authority of the pages and to re-rank retrieved pages. An additional gap has been over-stepped in event detection and even tracking activities (e.g. in the textual news) (Chang, 2001).
However, new tools have to be defined in order to adapt to the huge amount of heterogeneous information the organisations (governments, industries) have to take into account in their activities. Cross IM is an issue to be able to derive strategic information from various sources. More advanced customized tools should be possible thanks to IM.

2.5 E-commerce

Electronic commerce refers to the selling and buying of information, products and services using network facilities and more specifically via Internet. E-commerce requires to a large range of technologies to support the deep changes in this new way of business: electronic retailing, online advertising, online catalogues, interactive marketing, and electronic banking. One activity related to IM and IR is customized related. The goal is to profile customers by collecting information about consumers (user's interests, hobbies, and habits). This information is harvested either by asking the users them-selves or by spying on their connections, information they look at or actions on the system. This information on users is mined in order to create customer's profile and to provide them with customized advertisements, catalogues of products or services, interfaces.

2.6 Visualisation

The increasing quantity of data stored and available through networks presents challenges with regard to locating information, manipulating it and browsing it. New type of intelligent interfaces is needed in order to help users to know what the sources contain and to help him browsing these

sources and link the information contents. These interfaces should provide synthetic views of large sets of documents or pieces of information and should be able to focus on targeted information.

3 Methods and tools

IM can be paralleled to data mining that has been introduced in database technology through Knowledge Discovery in Databases (KDD). KDD intends to extract or infer useful and previously unknown knowledge from growing volumes of data (Fayyad, 1996) for knowledge induction and decision support starting from these masses of data. This goal is reached by discovering global patterns and relationships among the data. IM strives towards the same goal from less structured information. A knowledge discovery process consists traditionally of three stages: targeting the information to mine, mining it and interpreting it.

3.1 Knowledge discovery stages

Information selection and pre-treatment: This consists in collecting, homogenising, cleaning and reducing the data. This phase consists in improving the space of data to explore in order to provide the only useful pieces of information to mine according to the users' objectives. This is achieved by selecting the sources that traditionally support the information systems (the web being one among others resources), filtering and summarising the information they contain. The *filtering* is based on the sources querying and in extracting the objects or concepts to analyse. *Summarising* is used to turn the detailed data into more global data. The processes used depends on the mined sources (e.g. type of documents: texts, sounds, images, videos), on the mining applications and objectives.
Information analysis and mining: It consists in mining the cleaned information in order to discover existing but previously unknown relationships among the data.
Interpretation: The objective is to fulfil the user's needs in term of knowledge and to allow him or her to take the relevant decisions. The information produced by the analysis or mining process can either be used by a human user or by an application agent. Therefore, the results have to be presented in the most synthetic and expressive way.

3.2 Mining functions

The objective is to analyse the targeted information in order to discover existing but previously unknown relationships among the data. Mainly two means have been investigated in parallel: multidimensional analysis and rule discovery.
Multidimensional analysis is typically interactive and uses rollup and drill-down operations.
The dimensions are defined according to the type of information manipulated by the operational sources and answer the what? Who? Where? When? questions. As an example, in a technology monitoring activity, the dimensions can be the technology objects, the organisations that defined it, the time, the countries, etc. (Mothe, 2003). Generally, the dimensions are hierarchical so that rollup (increasing the level of aggregation, looking at more global information) and drill-down (decreasing the level of aggregation, looking at the detailed data) operations can be performed. To multidimensionality analysis is associated a graphical interface that allows to visualised the figures of 2 or 3 dimensions simultaneously.
With regard to *rule discovery*, several classes of mining functions have been defined:
- *Classification and clustering:* the objective is to find a partition of the data (mapping them into predefined classes or into clusters constructed according to the data similarities). Classification and clustering methods use a similarity measure to compare the different elements to classify.

- *Dependencies*: it includes the discovering of (weighted) dependencies between variables, relations between fields, temporal dependencies or sequences, regressions. One tries to find data correlation e.g. under the ('antecedent / consequence') form.

Mining functions use data analysis methods (Lebart, 98), natural language processing, machine learning, visualisation Human-Machine Interfaces.

3.3 Standards

Different *standards* are defined that will ease IM. Among them, we can quote (www.w3.org):

- MPEG7 (multimedia content description interface) that allows the description of multimedia data contents,
- XML (eXtensible Markup Language) that imposes constraints on the logical structure of documents and ease exchange of data on the web and elsewhere (interoperability),
- RDF (Resource Description Framework) that aims at modelling meta-data about the resources of the Web,
- DAML+OIL which is a semantic markup language for Web resources,
- Topic Maps which is "a language for exploring n-ary associations of members, each with its own role in the association".

These standards will play an important role not only in the semantic web but in new ways of representing, storing, searching and browsing information sources and contents as well.

4 Future

Because of the increasing growing of amount of information and its importance in nowadays society, there is an urgent need of a new generation of methods and tools. These tools should help the users in finding the nuggets that are hidden in these masses of multimedia data they potentially have access to. Mixing different resources (different media, different sources) and cross information and cross media mining should help users having precise ideas of a given field or providing precise answer to questions. IR challenges will covers not only efficient access to information that is stored but creation of new information from targeted heterogeneous data. Many domains will benefits of this kind of applications: health care, economy, and science.

5 References

T. Berners-Lee, J. Hendler, and O. Lassila, (2001), The Semantic Web, *Scientific American*.

G. Chang, M. Healey, J. McHugh, and J. Wang (2001), Mining the World Wide Web, An Information Search Approach, *Kluwer Academic Publishers*, ISBN 0-7923-7349-9.

L. Lebart, A. Salem, and L. Berry, (1998), Exploring Textual Data, *Kluwer Academic Publishers*, ISBN 0-7923-4840-0.

S. Chakrabarti, (2003), Mining the web, Discovering Knowledge from Hypertext Data, *Morgan Kaufmann Publishers*, ISBN 1-55860-754-4.

J. Mothe, C. Chrisment, J. Alaux, and B. Dousset, (to appear) DocCube: multi-dimensional visualisation and exploration of large document sets, *JASIST 'Web retrieval and mining'*.

Ontology, (2002), Different ways of representing the same concept, Communications of the ACM, 45(2).

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, (1996) Advances in Knowledge Discovery and Data Mining, *AAAI Press*, ISBN 0-262-56097-6.