Data Mining Techniques for Information Retrieval

Bikash Mukhopadhyay Sripati Mukhopadhyay

Abstract

Data mining automatically and exhaustively explores very large datasets, consequently uncovering otherwise hidden relationships among data. This technology has been successfully applied in science, health, marketing and finance to aid new discoveries and strengthen markets. In addition, data mining techniques are being applied to discover and organize information from the Web. Unfortunately these advancements in data storage and distribution technology have not been accompanied by respective research in data retrieval technology for a long time. To put it in short: we are now being flooded with data, yet we are starving for knowledge. This need has created an entirely new approach to data processing - the data mining, which concentrates on finding important trends and meta-information in huge amounts of raw data. In this paper the main concepts of data mining and automatic knowledge discovery in databases are presented (clustering, finding association rules, categorisation, statistical analysis).

Keywords : Data Mining, Text Mining, Web Searching, Natural Language Processing, Machine Learning

0. Introduction

More than five hundred years ago Johannes Guttenberg started the avalanche that has been recently given a name “information explosion”. Some researchers argue that information explosion phenomenon is mostly a psychological effect. For centuries people have been concocting - and even sometimes writing down - an innumerable amount of stories, tales, and philosophical and scientific findings. Human invention in producing the data (putting aside its usefulness) seems to be more or less constant. Only the recent advent of telecommunication systems and aforementioned dramatic drop in publication costs allowed people to suddenly realise how much information is actually being produced by humanity. This may be true, but as we are not going to revert to our previous unawareness for amount of available information, we must learn to navigate this new ocean of data. This can be a real barrier, as our navigational aids - library indices, search engines, software agents - are still very primitive and ineffective. Very often when we try to find a piece of information via the Web the search engines return thirty, fifty, even hundreds of “hits”. We suspect that information crucial for us is probably buried somewhere between the returned pages - but where exactly? In most cases it is impossible to read all these pages and assess them, throwing out irrelevant “hits”. Now the publishing channel has been greatly shortened, so our knowledge repositories (mainly the Internet) contain a lot of garbage data. How to distinguish it from valuable information? The information explosion emerged as the data storage and transfer technology achieved its maturity. Now researchers should concentrate more and more on devising new ways of dealing with such huge amount of data that would allow us to retrieve necessary information effectively, and to extract real knowledge from it. In short, we need methods for “distilling” the data.

1. Data mining - automatic knowledge discovery

This is well illustrated by a proverb popular among data mining community:

Although we have large sets of information at our disposal - we are still starving for knowledge.

But - what is knowledge? I will try to define that concept graphically :

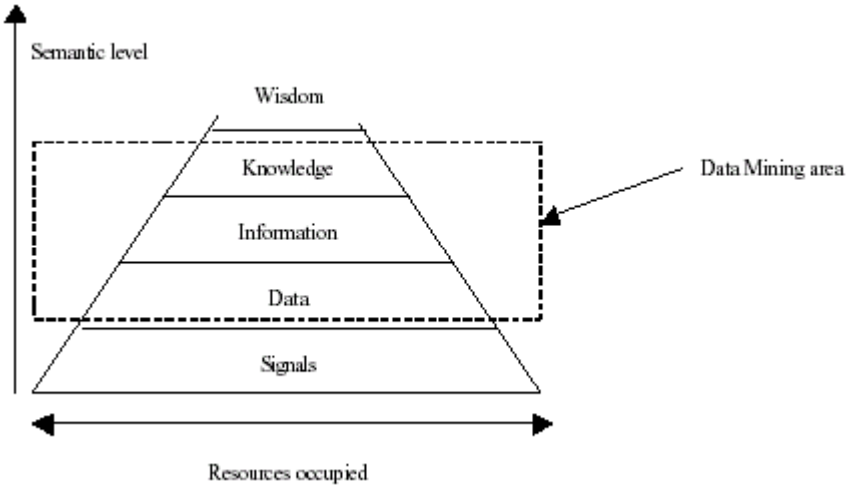


Figure: 1 Information pyramid

Let's explain above concepts using the telephone directory example. We would be dealing with such directory in electronic format, so one of lowest semantic levels would be bytes. These, together with ASCII code interpretation, represent strings of characters. For example we can encounter a sequence of bytes that after decoding would give us such sequence of characters: 6133560, what definitely can be regarded as some kind of data. This number could mean anything. It could represent number of people in the world wearing red jackets, but because we know that we are dealing with telephone directory, we can interpret that string as a telephone number, therefore jumping on the higher semantic level and obtaining a piece of information. Now we can start analysing further relationships between objects within this telephone directory. That can possibly lead us to discovery that 6133560 is Piotr Gawrysiak's, living in Warsaw Wawer district, telephone number. Moreover, further analysis shows that all telephone numbers that begin with digits 613 belong to person living in that district or in its proximity. This conclusion definitely has higher semantic importance than raw data analysed, and therefore we would classify it as a newly discovered knowledge. We of course do not know whether this particular piece of knowledge is useful. Even if it is - we have no idea how to use it. Such decisions do not seem to be amenable to computerisation and in fact could be called "wisdom".

Above example is very crude, but illustrates the point of Data Mining - using raw data, that *per se* does not have any visible underlying meaning, we extracted important semantic information. That information enriched our knowledge about the external world. Now we can define Data Mining more precisely:

Data Mining (DM) is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from large databases 6.

Two words are crucial in above definition: DM is an automatic process that - once tailored and started - can be run without human intervention (as opposed to OLAP), and databases that DM mines knowledge from are very large, and therefore not subject to human analysis. Data Mining is not a single method, or algorithm - it's rather a collection of various tools and approaches sharing the common purpose - to "torture the data until they confess". The results of Data Mining analysis can be miscellaneous, ranging from discovering customer behaviour, to fraud detection and automatic market segmentation, to full-text document analysis.

2. Main Methods of Data Mining

2.1 Association rules

Association rules finding is perhaps the most spectacular example of Data Mining, because it can quickly contribute to sales volume or profit when correctly implemented. Association models find items that occur together in a given event or record. They try to discover rules of the form:

if an event includes object A, then with certain probability7 object B is also part of that event.

Consider for example large supermarket network using association rules finding to analyse their databases. These databases contain information about transactions made by customers: articles bought, volume, transaction time etc. During the analysis process such hypothetical rules could be discovered:

If a male customer buys beer, then in 80% of cases he also buys potato chips

or

If a customer is paying at cash desks 1-5, then in 60% of cases he is not buying the daily newspaper.

Using these rules some strategic decisions could be made. The potato chips stand could be moved away from the beer stand, to force customers to visit more supermarket space. Special “beer plus chips” bundles could be introduced for customers’ convenience. The newspapers stand could be probably installed near cash desks 1-5 and so on.

2.2 Classification & Clustering

The data that we are dealing with is very rarely homogenous. In most cases it can be categorised using various criteria. For example company’s customers can be divided into various segments according to their weekly purchases volume, scientific texts can be divided by science discipline, and further into full papers and abstracts and so on.

The characteristics of such segments and their number provide us with substantial information about the nature of our data. Moreover even the sole fact that our data can be divided into different segments can be sometimes important. In data mining we distinguish two types of such segmentation process. First of them is classification, which is a learning process aimed at determining a function that maps - in other words classifies - a data object into one, or several, predefined classes. Classification employs a set of pre-classified examples to develop a model that can classify other records.

Clustering on the other hand maps a data object into one of several categorical classes but in this case they have to be determined from the examined data. Such data clusters that emerge during clustering process are defined by finding natural groupings of data items based on similarity metrics or probability density models. Classification and clustering is in classical data mining used most often for purely marketing purposes, such as market or competitors segmentation. These methods proved to be very useful in text mining (see section 5).

2.3 Statistical analysis

Statistical analysis is usually regarded as the most traditional method used in data mining. Indeed, many statistical methods used to build data models were known and used many years before the name

Data Mining has been invented. We must however remember that these simple techniques cannot be utilised in Data Mining without modifications, as they will have to be applied to much larger data sets than it is common in statistics. In effect a whole new breed of advanced artificial intelligence methods, combining conventional statistical tools with neural networks, rough sets and genetic algorithms has been recently created. The most widely used simple statistical method is regression. Regression builds models basing on existing values to forecast what other values, not present in input data set, could be. There are many possible applications of regression, the most obvious being product demand forecasts or simulation of natural phenomena. Three methods presented above are perhaps the most common tools used in data mining, mainly because they are especially good in dealing with numerical data. Extracting useful information from large amounts of textual information needs slightly different approach, what does not mean that experience gained from classical data mining research can not be reused there.

3. Full Text Documents Analysis

Full text document analysis is one of the most difficult problems in modern computer science, mainly because it is closely related to natural language processing and understanding. Processing of human language has proved to be much more challenging task, that it seemed in early sixties or seventies, and is still - as a technology - in it's infancy.

Fortunately a lot of problems related to "information explosion" can be coped with by using quite simple and even crude approaches, that do not need the computer system to *understand* the text being processed. Data Mining methods - like clustering and categorisation - can be effective here, because they don't rely on external information (such as extensive use of text semantics), and organise data using only relationships contained within it. Below I present a quick overview of most important problems related to full text document retrieval together with examples of solutions utilising data mining - like approaches.

3.1 Problems

Among all problems related to full text analysis two seem to be currently the most important. These are: poor quality of search engines - especially Internet search engines, and lack of automatic text categorisation tools which would allow for quick assessment of large document collections.

3.1.1 Internet search engines

Almost everyone agrees that current state of the art in Internet search engine technology means that extracting information from the Web is an art itself. Widely used search engines, such as [W2] and [W5] are plagued either by the lack of precision or by inadequate recall rate. They tend to return thousands of answers for even specific queries while from time to time refusing to find appropriate documents albeit they exist and are accessible through the net.

Almost all commercial search engines use classical keyword-based methods for information retrieval. That means that they try to match user specified pattern (i.e. query) to texts of all documents in their database, returning these documents that contain terms from the query. Such methods are quite effective for well-controlled collections - such as bibliographic CD-ROMs or handcrafted scientific information repositories. Unfortunately the Internet has not been created, but it rather evolved and therefore cannot be treated as well controlled collection. It contains a lot of garbage and redundant information and what is maybe even more important - it does not have any kind of underlying semantic structure, that could facilitate navigation.

Some of the above issues are result of improper query construction. The questions directed to search engines are often too generalised (like "water sources" or "capitals") and this produces millions of

returned documents. The texts that the user was interested in are probably among them, but cannot be separated as the human attention seems to be constant - one hundred documents is generally regarded as maximum amount of information that can be still useful in such situations.

On the other hand documents sometimes can not be retrieved because the specified pattern was not matched exactly. This can be caused by flexion in some languages, or by confusion introduced by synonyms and complex idiom structures (English word Mike is often given as an example of this, as it can be used as a male name or a shortened form of a noun “microphone”)8. Most search engines also have very poor user interfaces. The computer aided query construction systems are very rare, and search results presentation concentrates mostly on individual documents, not allowing for more general overview of retrieved data (which could be very important when number of returned documents is huge).

Last group of problems is created by the nature of information stored on the Internet. Search tools must deal not only with hypertext documents (in the form of WWW pages) but also with free-text repositories (message archives, e-books etc.), FTP and Usenet servers and with many sources of non-textual information such as audio, video and interactive content.

3.1.2 Text categorisation

It would be much easier to cope with “information explosion” and digest all data that is flooding us, if we could at least identify main subjects of all documents at our disposal, and further organise these subjects into some kind of structure, preferably hierarchical. A classical approach to this problem would involve building a handcrafted index and in fact such indices are in widespread use among the Internet [W4], [W5], and juridical communities. Unfortunately they simply cannot cope with the number of new documents created every day. It means that they tend to be more and more incomplete as the number of information available increases faster than index creators can analyse and classify it. Certainly, the need for automatic categorisation is really strong here.

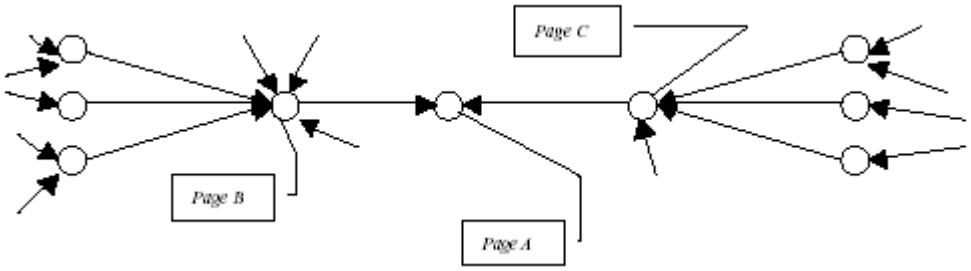
3.2 Solutions

I will not try to present all research results related to text mining here, as this would be an impossible task. Instead I will focus on innovative technologies developed especially with the Internet, or similar hyperlinked environment9, in mind. I am also not presenting here these new search methods, which do not have much in common with data mining. Such techniques include new generation of web page presentation tools [W7], autonomous software agents, and topic oriented search engines [W3]. Practically all new document retrieval and analysis methods fall into one of two groups. First of them includes techniques exploiting practically only hyperlink information and not being very concerned with actual text contents. This approach is possible because the hyperlinks are human-created entities, and therefore represent additional layer of semantic information, describing relations between document contents. Second group comprises of tools dealing only with raw text, and performing mainly some kind of statistical or associative analysis. These methods do not relay on hyperlinks and therefore have wider scope of possible applications.

3.2.3 Link-based methods

3.2.3.1 PageRank

As I already mentioned the hyperlink structure of the Web provides a lot of semantic information that can be used while assessing web page quality. The most obvious method, adopted from the bibliometrics field, would assign an authority index (or “weight”) to a page basing on number of hyperlinks (in other words “citations”) coming to this page. This method is simple and straightforward, but can be easily confused. Consider for example the following network, representing part of the worldwide web:



If a classical algorithm is used *Page A* would be assigned very low authority value as opposed to pages *B* and *C*.However, we intuitively know, that *Page A* could be important because it's relatively easy to get there using hyperlinks, from such different, not directly connected and widely cited parts of the Web as *Page B* and *Page C*. PageRank index has been conceived as solution to this problem. Its calculation simulates behaviour of so called "random surfer". Such hypothetical user starts browsing the Web from randomly selected page, and navigates it by clicking on the hyperlinks, writing down the addresses of visited pages. After certain amount of time (which is represented in this model as a number of "clicks") user gets bored, and starts anew from freshly selected random page. PageRank index value is defined by a probability that our random surfer visits given page.

Exact definition of PageRank is given below:

$$PR(A) = (1 – d) + d\{ PR(T_1) / C(T_1) + \ldots\ldots\ldots + PR(T_n) / C (T_n) \}$$

where PR(A) - PageRank of page A; C(A) - number of outlinks from A; d - simulates random surfer path length,

T - pages linking to A

Practical experiments have shown that in most cases strong correlation exists between PageRank index and human assigned "authority score" of a page. In other words, most valuable and trusted pages tend to have high PageRank indices. This allows for easy categorisation of Web pages and can especially effective in sorting search engine results.

3.2.3.2 HITS

Link structure has been also used for automatic identification of strongly interconnected web page clusters. Such emergent groups of pages often share the same topic, and can be treated as a kind of "Web community". First approach to automatic isolation of such Web thematic collections was J. Kleinberg's HITS algorithm, developed later into full-blown information retrieval system called CLEVER. One of the most important findings of Kleinberg was the concept of authority and hub pages. In classical bibliometrics the number of citations contained in a document is rarely seen as a significant contribution to this document importance10. However in the chaotic structure of the Internet such pages rich in outgoing hyperlinks act as important landmarks, providing tables of contents and "road directions" for surfers. Klienberg calls such pages with a name "hub". Accordingly, the pages containing mostly valuable information and therefore pointed by many pages are called "authorities". In HITS algorithm we repeatedly assign each page two weight values: an authority score, and hub score, defined as follows:

$a(p) = \Sigma_{q \rightarrow p} h(p)$ ← authority

$h(p) = \Sigma_{p \rightarrow q} a(q)$ ← hub

Of course above approach would not be very helpful in categorising entire Web contents, but it is quite effective with semantically restrained sets of pages. We can for example use it to quickly find most important pages within search engine results, filtering out the rubbish. This can lead to spectacular effects with very general queries (like “bicycles”, “aviation” etc.) as HITS algorithm tend to identify pages created by special-interest groups or indexes to web resources on a given topic.

3.2.4 Content analysis methods

3.2.4.1 Document similarity based classification

To perform effective object classification we must be able to compute some kind of distance metrics between them. Internet pages give us much more possibilities here than raw text documents, because when try to determine level of similarity in between hypertext documents we can use such formatting information like number of hyperlinks, frequency of viewing, depth of children and so on. Very interesting attempt to use this information in classification has been made by Peter Pirolli and James Pitkow from Xerox Palo Alto Research Center. They have tried to assign documents from Xerox intranet to one of the following classes: index, source index, reference, destination, content, head, personal home page and organisational home page13. The method used by them involved checking the strength of several page properties (such as its size, or number of hyperlinks) and using following table to perform classification:

Node type	Size	Inlinks	Outlinks	Children depth	Similarity to children	Request frequency	Entry point	Precision
Index	-		+					0.67
Source index	-		+				+	0.53
Reference	+	-	-	-				0.64
Destination	+	-	-	-			-	0.53
Head			+	+	+		+	0.70
Org. Home Page		+	+		+		+	0.30
Pers. home Page	>1kb <3 kb					-	-	0.51
Content	+	-	-	-				0.99

+ means that this property should be strong for given node type, - means this property should be weak

This analysis, accompanied by simple statistical comparisons between pages and topology computations resulted in categorisation with quite high precision. Almost all content pages has been classified correctly, and in practically all other cases more than 50% of all analysed pages has been assigned to the correct group. Note that this approach does not deal at all with semantic meaning of documents, which at first seems necessary to distinguish for example personal home page from content page (i.e. - page that actually delivers information, and not facilitates navigation).

3.2.4.2 Concurrence analysis

Other very promising method for computing relationship strength between lexical objects (not only documents, but also smaller entities such as paragraphs or even words) is Latent Semantic Analysis. The primary assumption of this method is that there exists some underlying structure in the pattern of word usage among documents, which can be discovered using statistical methods. Latent Semantic Analysis uses singular value decomposition over lexical objects concurrence matrix to discover relationships between words (or phrases etc.) that are appearing in similar contexts. Consider for example14 the following two sentences:

1) The U.S.S Nashville arrived in Colon harbour with 42 marines

2) With the warship in Colon harbour, the Colombian troops withdrew

Classical text analysis systems (that is - not equipped with thesaurus) will not be aware of semantic similarity of words “U.S.S Nashville” and warship. The LSA analysis is however able to capture this relationship, because both terms appear in similar context of words such as “*Colon*” and “*harbour*”. Latent Semantic Analysis can have many applications in text retrieval. The most interesting seem to be:

- automatic thesauri building and query expansion : as LSA is able to grasp the semantic relationship between lexical units, it could be used to build a thesaurus base frame. Of course very careful document selection (that is - documents fed to LSA algorithm) is necessary to ensure high quality of such thesauri, which anyway have to be rechecked by human experts afterwards.
- automatic document grouping and topographic text visualisation : similarity between documents calculated by LSA is a good distance measure than can be used in classical clustering algorithms to discover topic focused groups in a large collections of documents. Such techniques can be used for example in analysis of corporate email archives, or Internet newsgroups. Some companies (see [W4] and [W8]) are also experimenting with using these similarity metrics in construction of three dimensional maps, representing documents space.
- finding semantically similar documents (an text mining application of case based reasoning), like matching abstracts to full papers, or identifying examination frauds.

4. Conclusion

The amount of information available to us increased enormously, while the methods of retrieving that information remained relatively ineffective. The main source of difficulties in text retrieval research was natural language understanding barrier, which proved to be much more challenging than anyone had envisaged before. Fortunately it turned out that a lot of useful full-text analysis could be performed without a need to understand analysed text contents, in a way similar to emerging Data Mining techniques. Grouping and retrieval algorithms that have been roughly presented in this paper extract the underlying semantic information directly from the structure of analysed documents.

5. References

1. S. Brin, L.Page: “Anatomy of a large -scale hypertextual Web search engine”, WWW7 Conf. Proceedings, 1998
2. D. Gibson, J.Kleinberg, P.Raghavan: “Inferring Web communities from link topology”, Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, 1998

3. Daniel P. Dabney: „The Curse of Thamus: An Analysis of Full Text Legal Document Retrieval", American Association of Law Libraries, 1986
4. S.Chakrabati : "Experiments in topic distillation", IBM Almaden Research Center, 1998
5. Marcin Frelek, Piotr Gawrysiak, Henryk Rybiñski : " A method of retrieval in flexion-based language text databases", IIS'99 conference proceedings, 1999
6. P. Gawrysiak : "Information retrieval and the Internet", PWII Information Systems Institute Seminars, 1999
7. C. Westphal, T. Blaxton : "Data Mining Solutions", Wiley Computer Publishing, 1998
8. V. Dhar, R.Stein "Seven methods for transforming corporate data into business intelligence", Prentice Hall, 1997
9. "Designing the next generation of knowledge management centers", Bar-Ilan University, Department of Library and Information Studies, 1999
10. [W1] www.google.com
11. [W2] www.altavista.com
12. [W3] www.yahoo.com
13. [W4] www.polska.pl
14. [W5] www.infoseek.com

About Authors

Mr. Bikash Mukhopadhyay is Information Scientist of Burdwan University, Burdwan-713 104, India and holds MCA and pursuing Ph D. Earlier he worked with PCL ,TTTI, DCL and CCP. His area of interests are Content Management , Data Mining , Knowledge Based Computing Systems. He is a life member of B.L.A and IASLIC.
E-mail : buclib@satyam.net.in

Prof. Sripati Mukhopadhyay is Professor and founder Head of Department of Computer Science of Burdwan University, Burdwan-713 104, India . He holds M.Tech and Ph.D. His research areas are in the field of Artificial Intelligence ,Discrete Maths & Cryptography, Knowledge Base Computing System and Data Mining. He has many publications in international and national level and a life member of Computer Society of India.
E-mail : dgp_uvcompsc@sancharnet.in