ISSN:0975-9646

Data Mining: a Healthy Tool for Your Information Retrieval and Text Mining

Santosh Kumar Rath[1], Manobendu Kesari Jena[2],Tapaswini Nayak[1],Biswajit Bisoyee[1]

[1]Dept of Computer Science & Engineering
Gandhi Institute For Education & Technology
Bhubaneswar, Orissa, India

[2]Dept of Computer Science & Engineering
Majhigharianai Institute of Science & Technology
Rayagada, Orissa, India

Abstract- Data Warehousing and Data Mining are widely used by many industries like banking, insurance, healthcare, security and many others, however very little work has been done for Text-mining. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. In this paper we describe text mining as a truly interdisciplinary method drawing on information retrieval, machine learning, statistics, computational linguistics and especially data mining. We first give a short sketch of these methods and then define text mining in relation to them. Later sections survey state of the art approaches for the main analysis tasks preprocessing, classification, clustering, information extraction and visualization. The last section exemplifies text mining in the context of a number of successful applications. Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems Undeterred by the text explosion. It involves analyzing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the interest at present is coming from the biological sciences

1. INTRODUCTION

As computer networks become the backbones of science and economy enormous quantities of machine readable documents become available. There are estimates that 85% of business information lives in the form of text [TMS05]. Unfortunately, the usual logic-based programming paradigm has great difficulties in capturing the fuzzy and often ambiguous relations in text documents. Text mining aims at disclosing the concealed information by means of methods which on the one hand are able to cope with
the large number of words and structures in natural language and on the other hand allow handling vagueness, uncertainty and fuzziness. In this paper we describe text mining as a truly interdisciplinary method drawing on information retrieval, machine learning, statistics, computational linguistics and especially data mining. We first give a short sketch of these methods and then define text mining in relation to them. Later sections survey state of the art approaches for the main analysis tasks preprocessing, classification, clustering, information extraction and visualization. The last section exemplifies text mining in the context of a number of successful applications.

2. RELATED WORK:

Using Data Mining Methodology for text retrieval

Data Mining (DM) is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from large databases. 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 behavior, to fraud detection and automatic market segmentation, to full-text document analysis.

2. 1. Main methods of data mining for information retrievals

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 analyze 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. 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 utilized 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 cannot be reused there.

2. 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, which do not need the computer system to understand the text being processed. Data Mining methods - like clustering and categorization - can be effective here, because they don't rely on external information (such as extensive use of text semantics), and organize 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 utilizing data mining - like approaches.

3. 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 categorization tools which would allow for quick assessment of large document collections.

3.1. Internet search engines

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. The questions directed to search engines are often too generalized (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 cannot 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.

3.2. Text categorization

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 organize 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 [W5], [W6], 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 analyze and classify it. Certainly, the need for automatic categorization is really strong here.

4 .SOLUTIONS

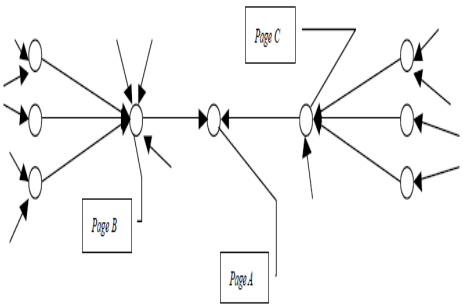
Practically all new document retrieval and analysis methods fall into one of two groups. First of them includes techniques exploiting practically only hyperlink information and not being very concerned with actual text contents. This approach is possible because the hyperlinks are human-created entities, 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.

Link-based methods

4.1. Page Rank

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 behavior 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. Page Rank index value is defined by a probability that our random surfer visits given page. Exact definition of Page Rank is given below:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right)$$

Where PR (A) - Page Rank of page A; C (A) - number of out links from A; d - simulates random surfer path length - pages linking to A. Practical experiments have shown that in most cases strong correlation exists between Page Rank index and human assigned "authority score" of a page. In other words, most valuable and trusted pages tend to have high Page Rank indices. This allows for easy categorization of Web pages and can especially effective in sorting search engine results. Practical implementation of such sorting method is currently tested in Stanford University [W1]. For detailed description of Page Rank calculations and its other possible applications see [3] and [4].

4.2. 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 importance.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".

$$a(p) := \sum_{q \to p} h(q) \leftarrow authority$$

$$h(p) := \sum_{p \to q} a(q) \leftarrow hub$$

Practical experiments show that after several iterations these weights seem to stabilize, thanks to mutually reinforcing hub-authority influence. The pages having highest authority or hub represent most important sources of knowledge and related hierarchical information and are closely interconnected. Of course above approach would not be very helpful in categorizing 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.

CONCLUSION

In this paper we discussed various methods such as page rank and Hyperlink-Induced Topic Search (HITS) and all limitations during the classification and retrieval of information. 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 analyzed 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 analyzed documents. Their simplicity and robustness give us hope that new generation of information retrieval tools will appear in near future.

REFERENCES

1.MacLennan Jamie, Tang ZhaoHui and Crivat Bogdan, □Data Mining with Microsoft SQL Server 2008□, Wiley India Edition.
2. G. Shmueli, N.R. Patel, P.C. Bruce, □Data Mining for Business Intelligence: Concepts,Techniques, and Applications in Microsoft Office Excel with XLMiner□, Wiley India.
3. Michael Berry and Gordon Linoff □Data Mining Techniques□, 2nd Edition Wiley Publications.
4. Alex Berson and Smith, □Data Mining and Data Warehousing and OLAP□, McGraw Hill Publication.
5. E. G. Mallach, □Decision Support and Data Warehouse Systems", Tata McGraw Hill.
6. Michael Berry and Gordon Linoff □Mastering Data Mining- Art & science of CRM□, Wiley Student Edition
7. Arijay Chaudhry & P. S. Deshpande, □Multidimensional Data Analysis and Data Mining Dreamtech Press
8. Vikram Pudi & Radha Krishna, □Data Mining□, Oxford Higher Education.
9. Chakrabarti, S., □Mining the Web: Discovering knowledge from hypertext data□,
10. M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis (ed.), □Fundamentals of Data Warehouses□, Springer-Verlag, 1999.