Vol-1, Issue-2 PP. 128-134 ISSN: 2394-5788

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

A Comparative Study on Text mining Techniques

Peerzada Hamid Ahmad
Research Scholar
M.M Institute of Computer Technology
& Business Management,
Maharishi Marakandeshwar University,
India.

Dr. Shilpa Dang
Assistant Professor
M.M Institute of Computer Technology
& Business Management,
Maharishi Marakandeshwar University,
India.

ABSTRACT

This research is intended to give detailed overview of basic concept of two text mining techniques namely information retrieval and information extraction. This paper has provided a comparison table of both these techniques on the basis of characteristic and their relationship to each other. We have also underlined many interesting research challenges which will benefit in managing the extracted valuable information. Later this research has also been motivated to highlight the main role of information extraction in terms of retrieval context in future to be played.

Keywords: Information retrieval, Information extraction, Text mining

1. INTRODUCTION

The volatile growth and attractiveness of the world-wide web has resulted in a vast amount of information sources on the Internet [1]. However, due to the heterogeneity and the lack of structure of information sources, access to this vast collection of information has been limited to internet browsing and searching. Text mining database are largely increasing due to increasing amount of information available in electronic form (electronic publication, documents, emails, conference material etc.). Nowadays most of the text is stored electronically in government sector, private organisation, industry, universities. This information which stored is in semi-structured or unstructured. For example, a document contains author name, publication date etc are in structured field but there are some fields which are unstructured like abstract, keywords and contents. Recently there has been great deal of studies on the modelling and implementation of this type of information [2]. Two techniques of text mining (information retrieval and information extraction) plays an important role in handling collection of information to trace a valuable information from this semi structured or unstructured text.

Information retrieval (IR) is finding documents of an unstructured text that satisfies information need from within large collections (usually stored on computers). Information retrieval is fast becoming the dominant form of information access [3]. Due to the rapid growth of text information, information retrieval system has found many applications such as on-line library systems, on-line document management systems, and the more recently developed Web search engines. Text mining techniques are used to extract this valuable information from the raw text data and then integrate to build a structured database. Information extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured text [4].
The rest of the paper is organized as follows. Section II introduces introduction, challenges and its role of a IR technique. Section III gives the detail about the IE technique, its role, challenges and its relationship with IR. Section IV gives the

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

comparison between IR and IE. Section V gives us future of IE in context to IR. Finally, the conclusions are made in Section 6.

2. INFORMATION RETRIEVAL

IR is the relatively old research area and gained increase attention with the growth of WWW. So there is a need for classy search engine. Information retrieval is the task of obtaining relevant information from a collection of resources [6]. It is used to focus on the textual information which includes text as well as document retrieval (web pages, pdf's, pp slides, paragraph's etc.). Document retrieval is measured as an extension of the information retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Each user tries to locate documents that can capitulate information that he or she requires i.e. each user tries to satisfy his or her information needs. The process of acquiring, identifying and searching the possible documents that may meet these information needs is called retrieval process [7]. All of the retrieved documents intend to satisfy user information needs expressed in natural language text. To studies the retrieval of information from a collection of written text documents is called Information retrieval (IR). Therefore information retrieval (IR) can be defined as a set of methods and techniques for formulating information needs of the users in form of queries. The query is then used to select a relevant document from a larger collection database (web) [8]. It can reduce information overload by using automated information retrieval systems. The information retrieval mainly deals with the large range of information processing from information retrieval to knowledge retrieval. Google search engine is the most well known information retrieval system which recognizes those documents on the WWW that are associated to a set of a given word [9]. This system is used by many universities, public libraries and companies to provide access to books, journals and other documents.
The most visible important IR application is web search engine in world wide web which recognize those documents on the WWW that are important to a set of a given words. The process of finding valuable information according to the user's request is information retrieval. It deals with indexing and retrieving of documents and also with crawling. Information retrieval system is used in online digital library, online service and online document system and web search engines. There are other powerful techniques in text mining like classification, categorization and summarization, clustering to handle large amount of text data [10].

Role of an IR system

- Exploring a problem domain, understanding its terminology, concepts and structure.
- Classifying, refining and formulating an information need
- Finding documents that match the information need description

Challenges for IR

There are a number of challenges associated with information retrieval system which can be solved by using existing search engines like Google [7].
Efficient way of

- ✓ Describing content method of documents
- ✓ Keyword matching from user in order to maximize the number of retrieved relevant documents.
- ✓ Storing information in a database.
- ✓ Eliminating the number of retrieved documents that are irrelevant and wrongly retrieved.
- ✓ Updating the information retrieval database with newly published web content.
- ✓ Designing and building large-scale retrieval systems is a challenging
- ✓ New retrieval techniques often require new systems

3. INFORMATION EXTRACTION

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

Information extraction (IE) is the task of automatically extracting structured specific information from unstructured or semi-structured natural language text. It identifies the extraction of entities such as names of persons, organisation, location and relationship between entities attributes events and relationships from text [11]. The valuable information extracted is without proper understanding of text such as name of a person, location and organisation [12]. These are stored in database like patterns and are then available for further use. In most of the cases this activity concerns processing human language texts by means of natural language processing. The information extraction process is accurate and robust access to particular information needs to be established. The information gathered is well-organized and stored automatically in a database. Its complexity of in use methods depends on the characteristics of source texts. The method can be rather easy and clear-cut if the source is well structured. If the source of information is less ordered or even plain natural language, the complexity of the extraction method becomes high as it includes natural language recognition and similar processes. The major advantage of information extraction systems is the precision of the queries and the clarity of the output, which can be efficiently reviewed, entered into a database or displayed visually. Main difference between IR and IE approaches is that in information extraction relevant facts of interest are specified in advance, while information retrieval tries to discover documents that may have facts of interest for the user that the user is not aware of. Information extraction is primarily based on pattern matching algorithms, so they rely on the structure of the information source [13].

Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction [14]. The main goal of information extraction is making information more accessible to the people and more machine-process able. From practical point of view, its main goal is to build large knowledge bases. There are three main problems associates with IE, namely paraphrase, ambiguity and data integration.

- Paraphrase- many ways to say the same thing.
- Ambiguity- the same word/ phase/ sentence may mean different things in different contents.
- Data integration which include the representation of an entity their relationship and large scale entity and relation resolution.

To automate the translation of text into structured data, a lot of efforts have been dedicated in the area of information extraction (IE). IE produces structured information ready for post-processing, which is vital to many applications of web mining and searching tools [15]. IE has useful variety of applications in different areas mainly given the recent explosion of internet and web documents [4]. It also includes application like medical patients records, whether forecasting report, conference announcements, advertising and announcement of jobs etc.

Main Functions performed by IE systems

- ✓ Expression analysis- Determines the expressions appearing in a document. This is especially useful for documents that contain many complex multi-word expressions, such as scientific research papers.
- ✓ Information extraction- Identifies and extracts complex information from documents which may be relationships between entities or events.
- ✓ Named-entity identification- Mainly specifies the names (people or organisation) appearing in a document. Some systems are also able to identify dates and expressions of time, quantities and associated units, percentages, and so on.
- ✓ IE transforms a amount of documents (textual) into a more structured folder, the folder assembled by an IE module then can be provided to the KDD module for promote mining of knowledge.

Challenges for IE

The main challenges for IE is given below

- ✓ Very large number of language (Natural) processing tools not available for all languages
- ✓ Much work on extracting more complex concepts like events, opinions, sentiments, entities, relationships.

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

- ✓ Disambiguating extracted mentions (Tracking mentions and entities over time) is common because text is inherently ambiguous; must disambiguate and merge extracted data
- ✓ Understanding, correcting, and maintaining extracted information like provenance and explanations, incorporating user feedback)
- ✓ Very difficult to build and maintain, very hard to port solutions across domains

IE is not IR

IE is not IR because of following two reasons:

- ➢ IE pulls facts and structured information from the content of large text collection (usually Corpora) while as IR pulls documents from large text collections (usually the Web) in response to specific keywords or queries.
- ➢ In IE, you analyze the facts while in IR you analyse the documents.

IE alternative to IR

IE as an alternative to IR because of following reasons given below:

- ➢ IE returns information at a much more deeper level than IR
- ➢ Database constructed through IE and connecting it back to the documents can provide a valuable alternative search tool.
- ➢ Even if outputs are not always perfect, they can be important if linked to the original text.

Relationship between IR and IE

IR and IE can be viewed as two different types of textual information access. It needs to understand documents in some manner, however the degree of understanding needed to accomplish each task is quite different. Based on differences (shown below in the table) these two tasks can be separated into two distinct categories. IR is the first category and IE is in the second category. The difficulty of textual information access increases from left to right as shown in the figure below.

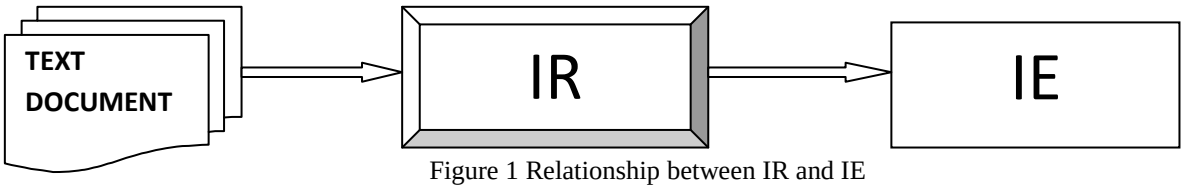
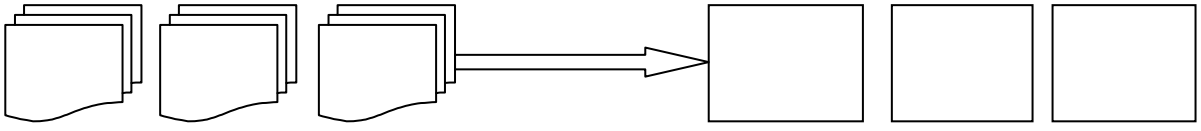


Figure 1 Relationship between IR and IE

4. COMPARISON OF IR AND IE

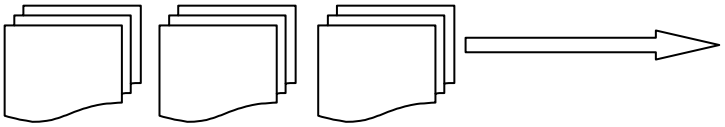
Information Retrieval (IR) is as the task of finding text documents, which are related to a user's information need. The Google information is the best IR system for the web where the result of an IR system is a subset of document that is related to a user's query. However, the goal of IE systems is to extract pre-specified features from documents rather than the documents themselves. The extracted features are usually entered into a database automatically. In short, IR is document retrieval while IE is feature retrieval. Figures 1 and 2 depict the difference between IE and IR.



GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

DOCUMENTS SUBSET OF DOCUMENTS

Figure 2 Information Retrieval



NAME	AGE	SEX
Rinku	20-25	F
Peter	35's	M

DOCUMENTS FEATURES FROM THE DOCUMENT

Figure 3 Information Extraction

Both IR and IE are difficult because they must overcome the ambiguities inherent in language. However, Information Extraction is often more difficult than Information Retrieval because IE requires more detailed knowledge about a document such as its organization, person, location, time, etc. Furthermore, IE systems are often required to establish relationships between features. IE is still very useful for applications in which users are only interested in specific features in a text rather than the whole text. For example, if a user only_wants to know names in newspapers, it is not necessary for a computer to understand entire newspapers. An IE system for name identification can work very well for this particular application. In general, Information Extraction is an easier problem than full natural language understanding since IE systems can ignore much of the information in text. The comparison [16] of two text mining techniques on the basis of characteristic is shown table 1.

Table 1 Difference between IR and IE on the basis of Characteristic

S. No	IR	IE
1.	Task of finding text documents which are relevant to a user's information need	Goal is to extract pre-specified features from documents or display information.
2	Document retrieval	Feature retrieval
3	Actual information buried inside document	Extract information from within the document.
4	Long listing of documents.	Aggregate over entire collection.
5	Describes details of Google which is best IR system for the web	Extracted features are usually entered into a database automatically.
6	Output of an IR system is a subset of documents that are relevant to user's query.	More difficult because it requires more detailed knowledge about a document such as its organisation, person, location, time etc. It often requires establishing relationships between features.

5. FUTURE OF INFORMATION EXTRACTION IN TERMS OF RETRIEVAL CONTEXT

From the above discussion whether it is relation or comparison we have seen that in future the IE and IR play an important role. Its main future role is given below:

- ✓ Multimedia Content Recognition- Videos, images, music etc.
- ✓ Cascaded Model-Output of one type of extraction forms the features of a more complex task of extraction.
- ✓ Queries related information recorded in a spoken format

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

- ✓ Information synthesis- satisfying or suppressing from valuable information
- ✓ Information integration
  - ➢ Additive or Sub additive accumulation- sum of information equal or smaller
  - ➢ Cooperative integration-sum of information larger (new information not explicitly present in the sources)
  - ➢ Disambiguation of one source of information with the other information

Today's world has become massively multimedia addiction. In addition to rapidly growing personal and industrial collections of music, photography and videos, media sharing sites have exploded in recent years. The growth of social media sites for not only social networking but for information sharing has further fuelled the broad and deep availability of media sources. Even special industrial collections are limited to proprietary access or precious books or esoteric scientific materials once restricted to special collection access, massive scientific collections or sensors once accessible only to few privileged users are increasingly becomes widely accessible.

An information synthesis is valuable information that is drawn from various resources. It depends on the ability to infer relationships among sources like essays, articles, fiction, and also non-written sources, such as lectures, interviews, observations. To draw relationships between two or more sources, you must understand the sources of information. It will frequently be helpful for the user. It also determines the relationship between the sources and which part should be used. It is the combination of useful information traced from various resources. Some relationships among sources must make them worth synthesizing. It follows that the better able you are to discover such relationships; the better able you will be to use your sources in writing syntheses and need to be more focused.

Information integration is the information which is provided to the user without giving background detailed instructions like how or from where to obtain the information. In the database, information from heterogeneous distributed information sources is gathered, mapped to a common structure and stored in a central location. In order to ensure that the information in the database reflects the current contents of the individual sources, it is necessary to periodically update the database. In the case of large information repositories, this is not feasible unless the individual information sources support mechanisms for detecting and retrieving changes in their contents. This is often an unreasonable expectation in the case of autonomous information sources. It has a drawback in the case of applications such as scientific discovery in which users often need to analyze the same data from multiple points of view: The database relies on a single common ontology for all users of the system. This ontology is typically specified as part of the database design. Each user queries the database using a common vocabulary and a common query interface which need to be focused.

6. CONCLUSION

In this paper we have described the two basic text mining techniques namely information retrieval and information extraction. During this study, the concept of both these techniques has been introduced and presented on the basis of characteristic. We have also highlighted some application and challenges but need to be more detailed focused on different areas in future use. There are many prospective research area in this filed to give better performance and accuracy in retrieving or extracting the valuable information from various resources. Combing a domain knowledge base with text mining engine would improve its efficiency, especially in the information retrieval and information extraction.

7. REFERENCES

1) Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled Shaalan, "A Survey of Web Information Extraction Systems", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, TKDE-0475-1104, 2000.

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

2) R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal Of Computational Engineering Research, ISSN 2250-3005(online), Vol. 2 Issue. 5, Page 1443, September 2012.

3) CD Maning, P Raghavan, H Schutze, "Introduction to information retrieval", Cambridge University Press, ISBN 978-0-521-86571-5, CUUS232/Manning, 2008.

4) Un Yong Nahm, Raymond J. Mooney, "Using Information Extraction to Aid the Discovery of Prediction Rules from Text", Appears in Proceedings of the KDD (Knowledge Discovery in Databases) Workshop on Text Mining, Page 51-58, August 2000.

5) AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, Ba-Quy Vuong, "Information Extraction Challenges in Managing Unstructured Data", SIGMOD Record, Vol. 37, No. 4, December 2008.

6) Ms.D.Subarani, "Concept Based Information Retrieval from Text Documents", IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Vol 2, Issue 4, , Page 38-48, July-Aug. 2012

7) R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", ACM, Press, Page 64, Year1999.

8) R. Gaizauskas, Y. Wilks, "Information extraction: Beyond Document Retrieval", Computational Linguistics and Chinese Language Processing, Vol. 3, No. 2, , Page 17-60, August 1998

9) R.Sagayam, S.Srinivasan, S.Roshini, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques". Internaltional Journal of Computational Engineering Research (ijceronline.com), Vol.2 Issue.5.

10) Varsha C. Pande1 and A.S. Khandelwal "A Survey Of Different Text Mining Techniques", IBMRD's Journal of Management & Research, Vol 3, No 1, Page 125-133, Year 2014.

11) K.L.Sumathy, M.Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues – An Overview", International Journal of Computer Applications, ISSN 0975 – 8887, Volume 80 – No.4, October 2013.

12) Vishal Gupta and Guruprit Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

13) M. Pejic Bach, N. Vlahovic, B. Knezevic, "Public Data Retrieval with Software Agents for Business Intelligence", In the Proceedings of the 5th WSEAS Int. Conf. on Applied Informatics and Communications, Malta, Pages 15-17, September, WSEAS Press, Athens, Greece, Page 215 –220, Year 2005.

14) Lokesh Kumar, Parul Kalra Bhatia, "TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS", Journal of Global Research in Computer Science ISSN-2229-371X, Volume 4, No. 3, March 2013.

15) Amit Kumar Mondal and Dipak Kumar Maji, "Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text", First Journal publication from SIMPLE groups, CLEAR Journal, Sept 2013.

16) Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, Pages 141-160, Sept 2003.

17) AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, Ba-Quy Vuong, "Information Extraction Challenges in Managing Unstructured Data", SIGMOD Record, Vol. 37, No. 4, December 2008.