Rough and Fuzzy Set Approaches to Information Retrieval Research: A Review

Vivek Badhe, Dr. R.S. Thakur, Dr. G.S. Thakur

***Abstract*— The majority of approaches to information mining focus on to improve information retrieves system performance. The main methodology underlying these approaches is to make approximate matches between query and documents in terms of recall level and ranking. Generally speaking, there are few approaches that have already been induced in modern search engines and crawlers. The purpose of this paper is to review the existing approaches of association rule mining, clustering and information retrieval based on rough set, fuzzy set and ontology that have been used to increase the effectiveness of information mining system.**

Index Terms— Information Retrieval, Rough Set, Fuzzy Set, Association Rule Mining, Clustering, Ontology

1. INTRODUCTION

Nowadays, large quantity of data is being accumulated in the data repository. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. However, in reality, a substantial portion of the available information is stored in text databases (or document databases), which consists of large collections of documents from various sources, such as news articles, books, digital libraries and Web pages. Since web search engines have become pervasive and search has become integrated, retrieving of information from these search engines consist of three essentials: query, documents, and search results.
The long history of information retrieval does not begin with the Internet. Information retrieval (IR) systems were found in commercial and intelligence applications as long ago as the 1960s. As with many computer technologies, the capabilities of retrieval systems grew with increases in processor speed and storage capacity. An IR system locates information that is relevant to a user's query (by using a set of keyword), typically searches in collections of unstructured or semi-structured data (e.g., web pages, documents, images, video, etc.), which might be useful to the specific purpose. The need for an IR system occurs when a collection reaches a size where traditional cataloguing techniques can no longer cope, which today is the relative case. With the growth of digitized unstructured information and, via high-speed networks, rapid global access to enormous quantities of that information, the only viable solution to finding relevant items from these large text databases was search, and IR systems became ubiquitous [5].
IR as a research discipline emerged with two important developments: how to index documents and how to retrieve them [5]. To index items by a list of keywords was a radical step in early days, and after a number of experiments use of words to index the documents of an IR system became established. With increasing data the need of its retrieval also arrived. To solve this so-called Boolean retrieval was used. Boolean model deals with using logical functions in the query to retrieve the required data. A query was a logical combination of terms which resulted in a set of those documents that exactly matched the query.
The concept of ranking was introduced to retrieve the result on the basis of *occurrence* of terms in the queries. This ranked retrieval approach to search was taken up by IR researchers, who over the following decades refined and revised the means by which documents were sorted in relation to a query [9].

Web Search

The Web has become the largest easy available repository of data. Hence, it is natural to extract information from it and Web search engines have become one of the most used tools in Internet. However, the exponential growth and the fast pace of change of the Web, makes really hard to retrieve all relevant information [5].

Data Challenges

Richardo Baeza-Yates [8] lists several data issues that need to be addressed. Among them some:

Dynamic Data:

The static Web has become small compared to content generated on demand, in particular by querying

Manuscript received July 07, 2014
Vivek Badhe, MANIT, Bhopal, India
Dr. R.S. Thakur, MANIT, Bhopal, India
Dr. G.S. Thakur, MANIT, Bhopal, India

in e-business or information services sites. Current search engines can follow dynamic links, but that has to be done with care, as there might be no limits or even the same page can be generated again and again. The queries asked can sometimes be more time consuming.

Multimedia Data:

Multimedia data includes images, animations, audio in several forms, and video. All of them have no standard formats. The ideal solution is to search on any kind of data, including text, using the same model and with a single query language.

Structured Data

Most data has some structure, leading to what is called semi-structured data. Examples are e-mail, news postings, etc. The first challenge is to design data models and associated query languages that allow mixing content and structure.

Semantic Data

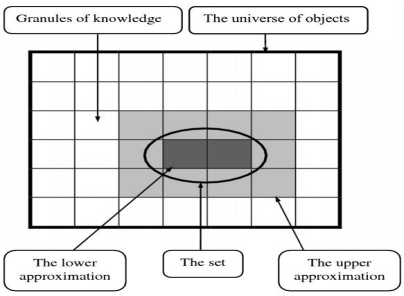
The two main problems with semantic information are standards for metadata that describe the semantic, and the quality or degree of trust of an information source. Other problems are common issues such as scaling, distributed authority, heterogeneous content and quality, multiple sources, etc.

2. Few Approaches to Information Mining

2.1 Rough sets: an introduction

Rough set theory, proposed by Pawlak in 1980s [6] can be seen as a new mathematical approach to deal with vagueness. The rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are similar in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory. In the rough set approach indiscernibility is defined relative to a given set of functional (attributes). Any set of all indiscernible (similar) objects is called an elementary set, and forms a basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as crisp (precise) set – otherwise the set is rough (imprecise, vague). Each rough set has boundary-line cases, i.e., objects which cannot with certainty be classified either as members of the set or of its complement, thus assuming that knowledge has granular structure. Due to the granularity of knowledge, some objects of interest cannot be discerned and appear as the same (or similar). As a consequence, vague concepts in contrast to precise concepts cannot be characterized in terms of information about their elements.

Therefore, we assume that any vague concept is replaced by a pair of precise concepts – called the *lower approximation* consisting of all objects which surely belong to the concept and the *upper approximation* containing all objects which possibly belong to the concept.



The difference between the upper and the lower approximation constitutes the boundary region of the vague concept. Approximations are two basic operations in rough set theory [1][6].

2.2 Ontology and Text Clustering

The concept of Ontology deals with various languages that are used for building semantic structures and increases its extraction efficiency. Semantic data and ontology is imagined as future web that uses text documents as well as semantic markup documents. We need to build a new paradigm for Information Retrieval (IR) that is compatible with all standards and provides effective and fast search. Thesaurus defines the set of standard terms or words that are used to search a document and set of relations between these terms. It gives hint related to words which are typed in search box. The selection of most discriminative characteristic of a linguistic unit that serves to distinguish it from other units of the same kind is employed by Ontology Label [3]. Ontology is very useful in defining the similar terms and with the help of clustering technique, the grouping of the words can be incorporated in IR. Semantic data selection technique improves the clustering that had already been performed using statistical selection method. Hence, motivated by co-occurrence terms and the semantic similarity between words, text clustering algorithm is used.

2.3 Fuzzy Set Theory

The notion of fuzzy sets provides a convenient tool for representing vague concepts by allowing partial memberships. A fuzzy set can be interpreted by a family of crisp sets, and fuzzy set operators can be defined using standard set operators. The membership values may be interpreted in terms of truth values of certain propositions, and fuzzy set operators in terms of logic connectives in many-valued logic. This provides a

formulation of fuzzy set theory based on many-valued logic

Comparative Study of Fuzzy Sets and Rough Sets

Y. Yao illustrates how fuzzy sets and rough sets are generalizations of classical set theory. Both fuzzy set and rough set are applied for modeling vagueness and uncertainty. The rough set theory takes into consideration the indiscernibility between objects. The indiscernibility is typically characterized by an equivalence relation. Rough sets are the results of approximating crisp sets using equivalence classes. The fuzzy set theory deals with the ill-definition of the boundary of a class through a continuous generalization of set characteristic functions. The indiscernibility between objects is not used in fuzzy set theory. A fuzzy set may be viewed as a class with unsharp boundaries, whereas a rough set is a crisp set which is coarsely described [7].

Recent Approach: Rough Set Theory to IR Field

As an extension of conventional set theory, it is considered as a good mathematical tool to deal with vagueness and uncertainty in data. Rough set theory analyzes data in an information system (e.g., information table). Objects of the universe are classified into equivalence classes if they share the same attribute values. For an arbitrary set of objects in the universe, we may not be able to represent it by using the equivalence classes induced by certain attribute set. The notion of approximation is introduced, in which two approximation operators, called lower approximation and upper approximation, are used to estimate the set of objects. Rough set theory has been used in many research fields, such as information retrieval, machine learning, and data mining. It is found to be particular useful for feature selection, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc [1] [6].

As an interdisciplinary field of study of text mining, the developments of information retrieval (IR) theories and techniques are accompanied by the growth of information science, cognitive science, statistics, linguistics and the World Wide Web. Since the first IR system was introduced in the 1950s [5], the searching contents, retrieval methodologies, and user interface of an IR system have been dramatically changed. In the late 1990s, the successful implementations of web search engines boosted many features of IR systems from experimental studies to World Wide Web applications, and consequently brought more challenges to information retrieval researchers. The more recent approach attempts to improve retrieval quality by using intelligent matching strategies. The rough set approach to information retrieval is following this direction.

3. Related Work

Rough set theory is a useful mathematical tool that deals with vagueness and uncertainty in data. It has been applied to many computer scientific fields, such as data mining, machine learning, pattern recognition, and expert systems [1][6][7].

One of the applications of Rough Set theory in machine learning is the so-called feature selection [2] especially for classification problems. This is performed by means of finding a reduce set of attributes which is a subset of all features which retains classification accuracy as original attributes. Finding a reduce set in decision systems is NP-head problem which has attracted many researchers to combine different methods with rough set.

Hong Zhou et. al [3] proposed an approach by using ontology and label mechanism which solve retrieval result information under the Chinese context, thus interacting information system at the layer of semantics and knowledge.

The explosive growth of information stored in unstructured texts created a great demand for new and powerful tools to acquire useful information, such as text mining. Document clustering is one of its the powerful methods and by which document retrieval, organization and summarization can be achieved [4]. However, it represents a challenge when dealing with a big number of data due to high dimensionality of the feature space and to the semantic correlation between features. In this paper, we propose a new sequential document clustering algorithm that uses a statistical and semantic feature selection methods. The semantic process was proposed to improve the frequency mechanism with the semantic relations of the text documents. The proposed algorithm selects iteratively relevant features and performs clustering until convergence.

Currently, most of the commercial information retrieval systems are based on the Boolean logic model [9]. They assume that a user's queries can precisely be characterized by the index terms. However, this assumption is inappropriate due to the fact that the user's queries may contain fuzziness. The reason for the fuzziness contained in the user's queries is that the user may not know much about the subject he/she is searching or may not be familiar with the information retrieval system. Therefore, the query specified by the user may not describe the information request properly. The fuzzy set theory can be used to deal with imprecision. Yih-Jen Horng et. al [10] proposed a fuzzy agglomerative hierarchical clustering algorithm for clustering documents and to get the document cluster more flexible and more intelligent than the existing methods due to the fact that it can expand users' queries.

4. Discussion and Future Scope

The majority of rough set approaches to information retrieval focus on applying rough approximation to improve IR system performance by approximate matches between query and documents. Based on the equivalence relation, documents that are similar to each other will be put into the same equivalence class. This strategy provides the maximum flexibility of what types of objects one can choose.

Search efficiency in the system of circulation of information, which ensures the expanded reproduction of knowledge, is associated with the choice or the introduction of special terminology. Surely, ontologies represent one of the promising areas in the formalization of knowledge, which enables their computer processing.

When using ontologies, there will be a need to group similar semantic data and documents which therefore can be retrieved with a good deal relative to the traditional retrieval methods. Clustering technique can be used to group similar semantic data and query through these clusters. Association rule mining can also be used to find the similar pattern behavior occurring in query relative to ontology, though being diverse in nature. It could help in bringing out rules that are more optimized and will retrieve more relevant data for the user.

There are certain drawbacks of rough set theory [6, 7] which could be overcome by fuzzy sets. Y. Yao confer that both the techniques have their certain merits either of which could be used in increasing the performance of information retrieval system in terms of document ranking, recall level and even more user oriented search strategies. Another purpose of the association mechanism in information retrieval systems using fuzzy set is to build the association relationships between index terms and to modify the user's queries by adding or replacing index terms associated with the queries which should find more relevant documents than that of the original user's queries and thus improve the retrieval performance [10]. Rough sets concepts can be pooled with fuzzy sets to find more appropriate fuzzy membership values that help in defining the granularity of knowledge. Therefore, the study of the association mechanism is very important in the field of information retrieval.

Moreover, other hybrid method such as machine learning can be used to enhance and increase the effectiveness of information retrieval systems. Thus using fuzzy set, rough set, clustering, association mechanism, etc could definitely be cooperative in decision making for an IR system.

5. Conclusion

The following conclusions can be deducted from the above review. The rough sets being capable of handling vagueness and fuzzy sets, which can deal with imprecision, could be used in combination with rough set to improve the query structure.

On the other hand, ontology is used to understand the query parameters by adding semantic knowledge to these terms. Clustering can be used in parallel to ontology by forming similar semantic terms clusters and than using the fuzzy-rough sets to improve the efficiency of query.

Acknowledgements

This work is supported by research project under Fast Track Scheme for Young Scientist from DST, New Delhi, India. Scheme 2011-12, No. SR/FTP/ETA-121/ 2011 (SERB), dated 18/12/2012

References

1. Bing Zhou, "Applying Rough Set Theory To Information Retrieval", 2013 26th IEEE Canadian Conference of Electrical And Computer Engineering (CCECE) © 2013 IEEE
2. Javad Rahimipour Anaraki, Mahdi Eftekhari, "Rough Set Based Feature Selection: A Review", 2013 5th Conference on Information and Knowledge Technology (IKT) ©2013 IEEE
3. Hong Zhoua, Bing-wu Liu, Jun Liu, "Research on Mechanism of the Information Retrieval Based on Ontology Label", 2012 International Workshop on Information and Electronics Engineering (IWIEE) 1877-7058 © 2011 Elsevier.
4. Asmaa Benghabrit, Brahim Ouhbi, Hicham Behja, Bouchra Frikh, "Text Clustering Using Statistical and Semantic Data" ©2013 IEEE
5. Mark Sanderson, W. Bruce Croft, "The History of Information Retrieval Research", Vol. 100, May 13th, 2012 © IEEE
6. Zdzislaw Pawlak, Andrzej Skowron, "Rudiments of rough sets", Information Sciences 177 (2007) © 2006 Elsevier Inc.
7. Yao, Y.Y., "A comparative study of fuzzy sets and rough sets, Information Sciences", Vol. 109, No. 1-4, pp. 227-242, 1998 © Elsevier Inc.
8. Ricardo Baeza-Yates, "Information retrieval in the Web: beyond current search engines", 2003 International Journal of Approximate Reasoning 34 (2003) 97–104 © 2003 Elsevier Inc.
9. Parul Kalra Bhatia, Tanya Mathur, Tanaya Gupta, "Survey Paper on Information Retrieval Algorithms and Personalized Information Retrieval Concept", International Journal of Computer Applications (IJCA) (0975 – 8887) Volume 66– No.6, March 2013.
10. Yih-Jen Horng, Shyi-Ming Chen, Senior Member, IEEE, Yu-Chuan Chang, and Chia-Hoang Lee, "New Method for Fuzzy Information Retrieval Based on Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques" IEEE Transactions On Fuzzy Systems, VOL. 13, NO. 2 2005