A FRAMEWORK FOR MULTILINGUAL TEXT REFINING IN WEB MINING

[1]Lalima Choudhary, [2]Bhavna Narain

[1]Research Scholar, [2]Associate Professor
[1] Department of Math's school of Information Technology,
[1] Mats University, Raipur, India

Abstract: The increasing volume of information available globally through the internet places high demands on information systems that can handle multilingual documents in a unified manner. Also, the languages used for Web documents are expanded from English to various languages. However, there are many unsolved problems in order to realize an information system which can handle such multilingual documents in a unified manner. The goal of this thesis is to provide some solutions to these problems. This paper proposed a framework by integrating techniques like cross-language information retrieval technique and sequence to sequence learning with neural network, which supports access to documents written in languages other than the user's native language. This framework provides some solutions to the problems in multilingual information processing that are specific to the Internet.

***IndexTerms* – Multilingual document, Web document, Integrating techniques.**

I. Introduction

From the user's point of view, three most fundamental text processing functions for the general use of the World Wide Web are display, input, and retrieval of the text. However, for languages some of the languages, character fonts and input methods that are necessary for displaying and inputting texts, are not always installed on the client side.

From the system's point of view, one of the most troublesome problems is that, many Web documents do not have meta information of the character coding system and the language used for the document itself, although character coding systems used for Web documents vary according to the language. It may result in troubles such as incorrect display on Web browsers, and inaccurate indexing on Web search engines.

Also, other text processing applications such as categorization, summarization, and machine translation are dependent on knowing the language of the text to be processed. Moreover, there might be some cases where the user wants to retrieve documents in unfamiliar languages, especially for cases where information written in a language other than the user's native language is rich. The needs for retrieving such information must not be small. Consequently, research on cross-language information retrieval (CLIR), which is a technique to retrieve documents written in one language using a query written in another language, are being paid much attention. However, it is difficult to achieve adequate retrieval effectiveness for Web documents in diverse languages and domains.

1.1 Text Mining

The data magnitude is growing gradually. About all type of associations, organizations, and business firms are storing their data by electronic means. A giant amount of content is flowing over the web in the form of digital libraries, stacks, repositories, and other contents such as email, chats, blogs, social media network [1]. It is challenging job to establish suitable patterns and trends to extort valuable facts from this large amount of data [2]. Conventional data mining tools are unable to maintain textual data since it requires time and attempt to extract information. Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources [3]. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics.

1.2 Multilingualism

Multilingualism and multiculturalism structures a vital part of the dynamics that describe today's universal information culture. The European Commission, (2007) offers the definition of multilingualism as: "the ability of societies, institutions, groups or individuals to engage on a regular basis with more than one language in their day to day lives." [4]. Consider to multilingualism in particulars, the definitions have turn out to be more inclusive and less firm. These descriptions tend to give emphasis to use as opposed to expertise: and describes, multilingual individual as "anyone who can communicate in more than one language, be it active (through speaking and writing) or passive (through listening and reading)" [5].

2. Literature Review

This section of the paper explores recent efforts and contributions on text mining techniques. It also deals with the important efforts in the field of cross language information retrieval (CLIR), natural language processing, and translation techniques.

2.1 A Comprehensive Study of Text Mining Approach

In this paper Abhishek Kaushik and Shudhanshu Naidhani basically describes a study on various mining approaches. Text mining or knowledge discovery is that sub process of data mining, which is widely being used to discover hidden patterns and significant information from the huge amount of unstructured written material. The propagation of clouds, investigations and technologies are accountable for the formation of huge amount of data. This sort of data cannot be used until or unless detailed information or pattern is discovered. For this text mining uses procedures of different fields like visualization, machine learning, text analysis, case-based reasoning, natural language processing, database technology statistics, and knowledge management [6].

2.2 A Study of Text Mining for Web Information Retrieval System from Textual Databases

K. Sankar, Dr. G. N. K. Suresh babu briefly describes the text information retrieval from textual database. This research describes the normal form of accumulated information is text, text Information Retrieval is believed to include a viable potential higher than that of data Information Retrieval. In truth, a recent study indicated that 85% of a company's information is contained in text documents[7].

2.3 Challenges and methods for multilingual text mining

In this research, Ralf Steinberger congregate insights by a variety of multilingual system developers on how to reduce the effort of developing natural language processing applications for numerous languages. Researcher also explains the main guidelines underlying our own effort to develop complex text mining software for tens of languages[8].

2.4 Natural Language Processing (Almost) from Scratch

In this study, researcher suggest a unified neural network structural design and learning algorithm that be able to be applied to a variety of natural language processing tasks including part-of-speech tagging, named entity recognition, chunking, and semantic role labeling. The benchmark tasks performed on this research is as follows –

1. Part-of-speech tagging
2. Chunking
3. Named entity recognition
4. Semantic role labeling [9].

2.5 Comparative Analysis of Data Mining Techniques, Tools and Machine Learning Algorithms for Efficient Data Analytics

This research examines a variety of free and open source data mining mechanism like WEKA, Rapid miner, Knime etc. The major is aim to uncover most precise tool and technique of classification practice. Relative analysis specifies that we can accomplish best result using a range of combinations of tools and classification technique. [10]

3. Proposed Approach: Methodology

The main of this paper is to develop a framework in which multilingual search content text (Hindi and English) will be input, for which we will use following methods and techniques. The Objectives is that it is necessary to build up text refining frameworks that process multilingual text documents and manufacture language-independent intermediate forms.

3.1 A Novel framework for multilingual search

A typical multilingual search engine assumes the existence of a query (which expresses the user's information need) and a set of documents (known as a document collection or corpus). These entities are route into internal representations appropriate for proficient comparison. The procedure of deriving document representations from a corpus is known as indexing. The term indexing engrosses extracting expressions, phrases, and concepts from the compilation and recording this information in a designed permitting rapid access. Just like this Query representations function in a much related fashion, although on a much smaller scale. By means of a diversity of different approaches, these representations are afterward compared to determine the "best fit." Documents that come out to match the query are then approved to the user, typically in the form of a ranked list. At this instant, the client is often given a prospect to respond to the subset of documents engendered by his/her query, providing feedback that can iteratively advance the results of the retrieval process.

By assessment, in cross-language information retrieval (CLIR), there is a linguistic discrepancy amid the queries that are submitted and the responses that are retrieved. To determine this inequality, CLIR engines are usually required to incorporate some facility for language translation, a clear necessity if query representations and document representations are to be significantly compared. Following are the common approaches to translation that can be employed at this point.

(1) Translate the query representation to match the document representations
(2) Translate the document representations to match the query representation

The approaches we will use are multilingual search utilizing dual translation.

3.2 Proposed Module

Following are the principle modules of our framework –

- Indexing
- Document representation
- Translation using machine learning through neural network
- Clustering
- Matching
- Ranked list of result

4. Result Analysis & Conclusion–

The region of search engine eminence research gains its significance not only from a broad attention of information science in the presentation of search systems but also from a wider debate on search engines in common culture. The expected outcome i.e. the novel multilingual search engine will be compared among some of the currently existing search engines on the basis of time taken for search i.e. Response time, precision in result as well as result description and the overall performance. Since the study used real users or potential users of Multi-Lingual Information Access tools, the study also highlighted practical application domains where Multilingual Information Retrieval (MLIR) technologies can be employed, thus helping motivate the need for further developments in MLIR while also providing an opportunity to evaluate the effectiveness of already existing technologies.

References

[1] SAVOY, J. 2007. “Why do successful search systems fail for some topics”, In Proceedings of the ACM Symposium on Applied Computing (SAC ’07). ACM, New York, PP. 872–877.

[2] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima, “Text Mining: Techniques, Applications and Issues”, (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7 No. 11, 2016.

[3] J. Vasuki, S. Priyadarshini, “ A study of basics of data mining, Machine Learning and Big data”, IJIRCCE, Vol. 5, Issue 1, January 2017.

[4] The Unicode Consortium. The Unicode Standard, Version 3.0. Addison-Wesley, Reading, MA, 2000.

[5] LIU, H., AND LIEBERMAN, H. Programmatic semantics for natural language interfaces. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI-2005) (Portland, OR, 2005).

[6] Abhishek Kaushik and Sudhanshu Naithan, “A Comprehensive Study of Text Mining Approach”, IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.2, February 2016.

[7] K. Sankar, Dr. G. N. K. Suresh babu, “A Study of Text Mining For Web Information Retrieval System From Textual Databases”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013.

[8] Ralf Steinberger, European Commission – Joint Research Centre (JRC) Via Fermi 2749, 21027, Ispra (VA), Italy, A survey of methods to ease the development of highly multilingual text mining applications.

[9] Ronan Collobert, jasonweston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, “Natural Language Processing (Almost) from Scratch”, Journal of Machine Learning Research 12 (2011) 2493-2537

[10] Lehtokangas, R., Airio, E., and Järvelin, K. (2004). “Transitive dictionary translation challenges direct dictionary translation in CLIR.” Information Processing and Management, 40(6), PP. 973–988.