Extracting Web Information for Clir

Archana M, Vijayalakshmi M. N, C. Chandrani, and Sumithra Devi K. A.

***Abstract*—The increasing need for the access to information without language barrier has lead to the strong demand for cross language information retrieval (CLIR). The concept of CLIR allows retrieving information written in different language from the language of the user query. One of the approaches to CLIR is based on the use of Web mining. Web mining use data mining techniques to automatically discover and extract information from web documents and services. In this paper we study the need of techniques that are useful for information search and cross language information retrieval. The advanced methods for acquiring the information using web mining techniques are elaborated, which improves the efficiency and effectiveness for cross-language web retrieval.**

***Index Terms*—Web mining, information retrieval, cross language information retrieval, translation.**

I. Introduction

The coverage of web information is very wide and diverse. All the documents in this data repository are not arranged in any order. Since web is highly dynamic information source, it is very difficult to retrieve the information. With respect to users (from different background, interest and usage purpose) point of view web information is updated frequently. It is observed that only the small portion of information posted on the web is useful.

Web mining is the application of data mining techniques to discover patterns from the web. Web data contains different kinds of information, including web documents data, web structure data, web log data, and user profiles data. Two different approaches can be proposed based on the definition of web mining:

- Process-based
- Data-based

In this reference, web mining can be defined as an application of data mining techniques to extract knowledge from web data, where atleast one of structure or usage data is used in the mining process. There is no much difference between web mining and data mining when compared. All web data can be mined in three different dimensions:

- Web content mining
- Web structure mining
- Web usage mining.

Web data are those that can be collectively used in the context of cross language information retrieval.

Web content data are used to design the web pages to be presented to the users. Web content data consists of free text, semi-structured data like HTML pages and structured data like automatically generated HTML pages, XML files or data in tables related to web content and also the data included in this category are textual, image, audio and video.

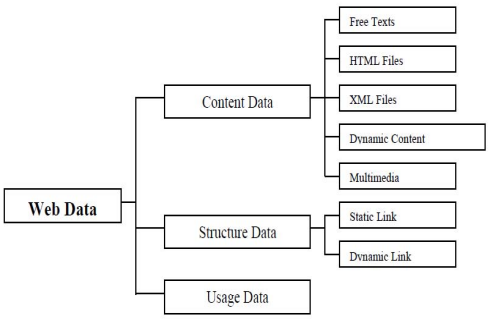

Fig . 1. Types of web data [1]

Web Structure data is used in the organization of the content. In a given page the arrangement of various HTML and XML tags gives the information about the intra-page structure. It is the hyper-links connecting one page to another. Web graph is constructed by hyperlinks information from web pages. Web graph is described as the core of web structure and used to represent web structure related to web page connectivity (dynamic and static links).

Web Usage data consist of web log data which are created on the web server such as web server access logs, proxy server logs, browser logs, registration data, cookies and any other data generated as a results of web user interactions with web servers. A unique IP address and a domain name are given to the Web server, whenever a user places the request. Web server fetches the page and sends it to user's browser. Web server data are created from the relationship between web user's interaction with a web site and the web server.

II. Multilingual Information Retrieval

Information Retrieval (IR) is to find relevant documents from a large collection of documents or from the World Wide Web. This can be obtained usually by formulating the user's query, often in free text, to describe the information need. The IR system then compares the query with each document in order to evaluate its similarity (or probability of relevance) to the query. The results, i.e. the retrieved documents are arranged in the decreasing order of their similarity. IR helps in getting useful information from digital resources including digital libraries, WWW and documents. The major problem in the IR System is that how effectively it has retrieved the relevant documents and rejected the irrelevant documents. [2]

Multilingual is also known as Cross-language information retrieval (CLIR), is a process where the user presents

Manuscript received September 10, 2011; revised December 15, 2011.

Authors are with are with the Department of Mater of Computer Application, R.V.College of Engineering, Bangalore, India (e-mail: archana.s.uday@gmail.com).

queries in one language to retrieve documents in another language. It is quickly becoming a mature area in the information retrieval world. The main goal is to allow a user to issue a query in language L and have that query retrieve documents in language L`. The barrier of the CLIR is what should be translated:

- Query may be translated
- Document may be translated
- Both query and the document may be translated

Knowledge Creation (Data mining) is a process of automatically discovering useful information in the large data set. Data mining techniques are deployed to search large databases in order to find useful patterns. There are four fundamental approaches for knowledge-based CLIR: cognate matching, document translation, interlingua techniques and query translation. [3]

III. Various Approaches for Multilingual

The basic approaches to CLIR are Machine translation, controlled vocabulary, dictionary-based and ontology approach

A. Machine Translation

Machine translation (MT) is considered as one of the most common method in CLIR, in which query or document is translated automatically. It is treated as the simplest form because the query given in one language can be translated into different language for search and the searched information can be translated back to the original one for viewing. It can produce high quality translations for a particular domain that contain some specific technical terms.

B. Controlled Vocabulary

Controlled vocabulary is considered as one of the most emphatic and effective approach. In this approach an interpretive layer of semantics is inserted between the term entered by the user and the underlying database for the better representation of the user's terms. A multilingual thesaurus is created to hold a list of descriptors for each document that has been retrieved and each term in the thesaurus must be translated for the languages involved. From the previous indexing, if the system is familiar with the most likely to important terms, then a descriptor can be added manually or automatically to the thesaurus.

C. Dictionary-Based

In dictionary based CLIR each term in the user query is looked up in the machine-readable bilingual dictionary. An equivalent selection method is applied to pick the best translation of that term from the list given by the dictionary and then added to the document language semantic mapping. It is then matched with document collected as though it was directly derived from the initial user request. Dictionary based method for CLIR can be put into four logical steps [4]:

- Pre translation query modification
- Dictionary lookup
- Equivalent selection and weighting
- Post-translation query translation modification

The advantage of using a simple bilingual dictionary to translate is that, it contain wordlists covering a wide range of subject areas and as well as language pairs are readily available. In addition, the time needed for set up a dictionary-based system from a printed or electronic source is less comparative to an MT engine for a new language pair. A machine-readable bilingual dictionary can be considered as a data structure that contains a list of dictionary entries for a given set of terms, and a lookup mechanism which, given a source-language query term, look upon this data structure to obtain possible translations or equivalents of the term in question. An entry in a machine-readable bilingual dictionary is a data structure within a dictionary containing all of the necessary information for a given spelling of a source-language query term

D. Ontology

Ontology is one of the basic approaches, in which domain-specific method is particularly used for the CLIR. In the domain-specific ontology relevant terms contained in a query are translated into several languages using the term-to-concept links established in the multilingual thesaurus. It is also valuable in the presentation generation process because, effective presentations are those that succeed in conveying the relevant domain semantic to the user.[5]

In comparison with machine translation or dictionary look-up methods for query and document it is found that dictionary-based query translation seems to work best for short queries while for long queries machine translation of the queries performs better than dictionary lookup. An important problem in the translation of short queries is the lack of context for the disambiguation of words that have more than one meaning and therefore may correspond to more than one translation.

The snag with controlled Vocabulary is that query can be generated using the vocabulary included in the thesaurus only, in this case it may be difficult to search for specific terms that is not present in the thesaurus. Its effectiveness is decreased as the number of vocabulary in the thesaurus increases.

Ontology generation is often haphazard depending on the skill of the user generating the ontology, it is also difficult to remain true to a guiding theory or philosophy while grounding ontology in what is happening in the world.

IV. Web Mining

The Web stands today as the world´s largest source of public information. Its magnitude can also be perceived as a drawback in a certain sense, however, nowadays there is a generalized problem in retrieving documents that may be written in any language, but through queries expressed in a single source language. And although Information Retrieval (IR) depends on the availability of digital collections, this key aspect is no longer the only concern. It is time for the multicultural society of Internet to make use of new technologies such as CLIR. Information retrieval systems' ability to retrieve highly relevant documents has become more and more important in the age of extremely large collections, such as the World Wide Web (WWW).

Web mining can be disintegrated as the following [6]:

- Resource Finding: task of retrieving the proposed web documents.
- Information Selection and Pre-Processing: the information retrieved from the web is automatically selected and pre-processed.

- Generalization: general pattern can be uncovered for a single or multiple web sites.
- Analysis: verification and validation of mined patterns.

There are three areas of Web mining according to the usage of the Web data as input namely: Web Content Mining, Web Structure Mining and Web usage mining.

A. Web-Content Mining

Web content mining focuses on retrieval of useful information from web content documents. The contributions of web content mining can be evaluated on two fronts: search result mining and web page content mining.

One of the tasks performed by the search engine is the search result mining, which retrieves information from the web and its utility relay on its capacity to retrieve the relevant document from a large collection of web pages centered on the users query. The most frequently used technique in search result mining is web document classification. Using this technique web document is consigned a keywords of varying degrees of confidence. A document classified in this way will be easy to be retrieved depending on its keyword-based searches [6].

The pages retrieved in this form are arranged according to its rank i.e. most relevant page appears on the top of the list of pages retrieved. A variety of approaches are used by different search engines to rank the web pages. An alternative method for information retrieval is the document clustering in which retrieved documents are partitioned into clusters depending on the topical homogeneity of web document.

Information Extraction (IE) is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts. Information is appealed from the texts of heterogeneous formats which include: e-mails, web pages, PDF files and organized into a single homogenous form. Web content is converted into the database that can be accessed by the query for the translation using any of the methods.[7]

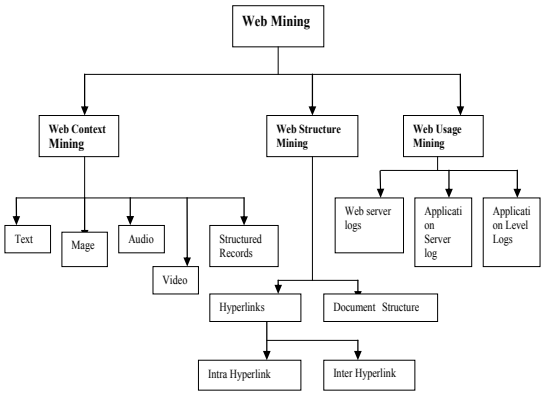

Fig. 2. Web mining TAXONOMY [8]

B. Web-Structure Mining

Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlink with or without the link description. This model can be used to categorize the Web pages and is useful to generate information such as similarity and relationships between Web sites [9]. Web structure mining can be considered as the process of learning structure from the web. This type of mining can be divided in to two type depends on the structure of data[8]:

- *Hyperlinks:* It is used for connecting one page to other or within the same page. When a hyperlink used for connecting different part of the same page is known as *Intra-Document Hyperlink* and the link which connect two different pages known as *Inter –Document Link.*
- *Document Structure:* The contents in web pages based on the *various* HTML and XML tags can be organized in a tree-structured.

Two algorithms that have been proposed are HITS and Page Rank. Both approaches focus on the link structure of the Web to find the importance of the Web pages.

- For a given query, HITS will find authorities (Pages *with* good sources of content) and hubs (Pages with good sources of links). "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs" [10].
- Page rank algorithm is used to calculate the importance of web pages using the link structure of the web. The Page Rank algorithm is defined as: "We *assume* page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:[11]
- $$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

C. Web-Usage Mining

Web usage mining is used to predict the behavior of the user in a web domain. It can be divided into different phases depending on the type of web usage data.

- Pre-Processing: the data retrieved from the web site tends to be incomplete, noisy and inconsistent. They should be shaped according to the requirement of the next phase, which includes the process of data cleaning, data integration, data transformation and reduction.
- Pattern Discovery: in this stage, user patterns (query) are identified with the help of the algorithms of statistics, machine learning and pattern recognition
- Pattern Analysis: the given pattern is interpreted, visualized and understood.

V. CHALLENGES OF WEB MINING

A. Web-Content Mining

Issues of web content mining in context of web warehouse can be discussed as:

- Finding the difference and the similarities between web content mining and conventional data mining.
- Selection of useful information from the web for web content mining before analysis
- To clean the selected data for effective mining
- Discover the different types of knowledge
- Collect the hidden information for decision making
- To make web content interactive

B. Web-Structure Mining

Structural information can be generated for the web tuples stored in web table by measuring the frequencies:

- For the local links in the web tuples. Local links connect the different web documents residing in the same server.
- For web tuples containing links which are interior; links which are within the same document.

- For web tuples, contains links that are global; links which span different web sites.
- For identical web tuples that appear.

C. Web-Usage Mining

According to [12] one of the approach is using maximal forward references that can be obtained by filtering out backward references from traversal subsequences in log data to extract frequently occurring consecutive subsequences that leads to maximal reference sequence, which are those frequent subsequences that are not subset of others.

VI. Conclusion

Due to the enormous growth in the usage of web and its associated technologies, the need for extracting useful information from web is also growing. Web mining techniques are also used to improve the effectiveness of CLIR. The following table gives the different approaches used by web mining under various categories:

TABLE 1: Approaches for Web Mining

Web content mining	Is the process of Information discovery from different sources of World wide Web structure
Web structure mining	is the process is used to categorize the Web pages
Web usage mining	is the process of mining for user browsing and access patterns and concludes by listing research issues

Therefore web mining can be used as an effective technology to retrieve information and improve the effectiveness of CLIR.

References

[1] M. A. Bayir, I. H. Toroslu, and A. Cosar, "A NewApproach to Reactive Web Usage Data Processing," *WIRI06, the 2nd International Workshop on Challenges inWeb Information Retrieval and Integration*, ICDE 06'sWorkshop 2006.

[2] W. Kraaij, J.-Y. Nie, and M. Simard "Embedding Web-Based Statistical. Translation Models in Cross-Language Information Retrieval."

[3] D. Soergel," Multilingual thesauri on Cross-LanguageText and Speech Retrieval," American Association forArtificial Intelligence 1997.

[4] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," In *Proceedings of the 20th ACM SIGIR conference on research and development in Information retrieval,* pp. 84–91, 1997.

[5] J. W. Lamp and S. K. Milton, "Grounded theory as Foundations for methods in applied ontology"

[6] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *ACM SIGKDD Explorations Newsletter*, Vol. 2 Issue 1, June 2000.

[7] K. C. Adams, "The Web as a database: New extraction technologies and content management," vol. 25, pp. 27-32.2001.

[8] J. Srivastava, P. Deskin, and V. Kumar, "Web Miming – Accomplishment & future Directions," tut-paper.pdf,Page no -51-70.

[9] K. N. Eiron, Mc Curley, and J. Tomlin, "Ranking the web frontier," in *Proceedings of the international conference on World Wide Web*, (WWW'04), pp. 309-318, 2004.

[10] J. M. Kleinberg "Authoritative sources in a hyperlinked environment". In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms,* pp. 668-677, 1998.

[11] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," in *Proceedings of the Seventh International World Wide WebConference*, 1998.

[12] S. Madria, S. S. Bhawmick, W.-K. NG, and E. P. LIM,"Research Issues in Web Data Mining,Center for advanced information systems," school of applied science, nanyang technological university, Singapore.PDF

[13] M. Archana, M. N. Vijayalakshmi, C. Chandrani, and Sumithra Devi K A, "Mining the Web Information ForCross Language Information Retrival," proceeding of ICMET2011.

[14] Y**.** Cao and H. Li, "Base Noun Phrase Translation Using Web Data and the EM Algorithm," *Proceedings of the 19th International Conference on Computational Linguistics*, 127-133, 2002.

[15] D. W. Oard, "A comparative study of query and document translation or cross-language information retrieval," *Third Conference of the Association for Machine Translation in the Americas, AMTA, 472–8 3,* 1998.

[16] K. L. Kwok, NTCIR-2 Chinese, "Cross Language Retrieval Experiments Using PIRCS," *Proceedings of NTCIR workshop meeting*, pp. 111-118. 2001.

[17] J. Y. Nie, P. Isabelle, M. Simard, and R. Durand, "Crosslanguage\ Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web," *Proceedings of ACM-SIGIR Conference*, pp. 74-81, 1999.

[18] C. Lu, Y. Xu, and S. Geva. "Web-Based Query Translation for English-Chinese CLIR," *The Association for Computational linguistics and Chinese Language Processing*, 2008.

[19] M. Archana, and Sumithra Devi K A, A Noval Approach For Classification Techniques for CLIR," "*Proceedings of the First International Conference on Computer Science*,Engineering and Applications Page no-246, CCSEA 2011

Ms.Archana.M, is presently working as Assistant Professor in Department of Mater of Computer Application, R.V.College of Engineering, Bangalore, India, She has completed her MCA from Gulbarga University, MPhil from Alagappa University and has 6 years of teaching experiences. She has published papers in national and international Conferences and Journals. She is also a member of IAENG.

Dr. Vijayalakshmi M.N. is presently working as Associate Professor in Department of Master of Computer Application, R.V.College of Engineering, Bangalore, India.She had completed her PhD from Mother Teresa Women's university, Kodaikanal in 2010. She has 12 years of teaching experience and 5 years of Research experience. Her research interests are Pattern recognition, data mining, neural networks, Image Processing. She is a life member of ISTE, CSI, and IACSIT.

Miss.Chandrani Chakravorty is presently working as Assistant Professor in Department of Mater of Computer Application, R.V.College of Engineering, Bangalore, India, she has completed her MCA and Mtech (Computer Cognition Technology) from Mysore university and has 4 years of teaching experiences. She has published papers in national and international Conferences and journals. She is the member of CSI

Dr.Sumithra Devi K. A, is serving as a Director of Master of Computer Applications Department at R.V.Collage of Engineering, Bangalore, India, she earned PhD in Computer Science and Engineering from the Avinashilingam University for Women, Coimbatore, INDIA and she draws a strong back ground in VLSI Partitioning CAD Tool. She has been awarded by many national and international awards. She is the lifetime member of IEEE, WIE, ISTE and CSI. She is a registered PhD guide under VTU and is guiding two research students and one research candidate under the funded project.