VOL. 10, NO. 7, APRIL 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

WEB PREDICTION METHOD ON SOCIAL NETWORK ANALYSIS

M. Mohamed Iqbal Mansur[1], C. Kavitha[2] and K. Thangadurai[1]
[1]Government Arts College, Karur, India
[2]Thiruvalluvar Government Arts College, Rasipuram, India
E-Mail: mansur.iqbal75@gmail.com

ABSTRACT

Web Mining is the integration of information gathered by traditional Data Mining methodologies and techniques with information gathered over the World Wide Web. The World Wide Web today provides users access to extremely large number of Web sites many of which contain information of education and commercial values. Web mining research, in its last 15 years, has on the other hand made significant progress in categorizing and extracting content from the Web. Nowadays the Web has proved to be as a rich and extraordinary data source of information, where multiple domains can be accessed and mined. Mining Web data is referred as Web Mining. Some of the objectives of mining web data include finding relevant information discovering new knowledge from web personalized, web synthesis and learning about individual users. Amongst these the most common use is finding relevant information. In this paper, we represent Web Prediction Method on Social Network Analysis. These techniques can be used for real world applications like market strategies, business intelligence and etc... Social Networks the interest of a single user represents the interest of the whole group. Ontology defines a set of representational primitives with which to model a domain of knowledge or discourse.

Keywords: web mining, data mining, world wide web, social networks, ontology.

1. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [2].

Web mining taxonomy

Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined.

A. Web content mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

B. Web structure mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web.

C. Web usage mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site.

Figure-1. Taxonomy of web mining.

2. EASE OF USE

Four Steps in Content Web Mining

When extracting Web content information using web mining, there are four typical steps.

- Collect - fetch the content from the Web
- Parse - extract usable data from formatted data (HTML, PDF, etc)
- Analyze - tokenize, rate, classify, cluster, filter, sort, etc.

- Produce - turn the results of analysis into something useful (report, search index, etc)

User profiles

The web has taken user profiling to new levels. For example, in a "brick-and-mortar" store, data collection happens only at the checkout counter, usually called the "point-of-sale." This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every action taken by the user, providing a much more detailed insight into the decision making process [1].

3. ONTOLOGY BUILDING FROM WEB

Ontology learning is defined as an approach of ontology building from knowledge sources using a set of machine learning techniques and knowledge acquisition methods. Ontology from texts is a specific case of Ontology from Web and has been widely used in the community f engineering knowledge since texts are semantically richer than the other data source type. These approaches are generally based on the use of textual corpora. This one should be a representative of the domain for what we are trying to build ontology. By applying a set of text mining techniques, granular ontology is enriched with concepts and relationships discovered from textual data. In such approach, human intervention is required to validate the relevance of learned concepts and relationships. In the last decade, with the enormous growth of Web information, Web has become as important source of information for knowledge acquisition: due to its huge size and heterogeneity. This has been the cause of mainly two categories of Ontology approaches: ontology learning from textual content of the Web, ontology learning from online Web ontology's, from web dictionary and from Web heterogeneous sources.

A. Ontology learning approaches from Web documents

Ontology from Web documents require the same techniques used before for ontology extraction from texts. Several approaches are based on eliminating tags from documents to obtain plain texts on which traditional text mining texts could be applied. We propose to classify these approaches to domain- dependant Ontology and incremental Ontology.

B. Domain- Dependent approach for Ontology learning from textual documents

Ontology approaches from Web content consists generally in enriching a small ontology called "minimal" or "granular" with new concepts and new relationships using text mining techniques. Learning ontology's from texts has been widely used in the community of knowledge engineering. This is in particular the work of: [9, 10, 11, 14, 15, 16, 17, 18]. However, no sufficiently detailed methodology has been presented to assist the learning process ontology. Indeed, the literature is limited to the presentation of guidelines more or less general. Thus, for each approach, it is important to know the aims and scope of the learning process, its main stages, the sources of knowledge used in learning, the main techniques applied in the process, re-usability of ontology's existing and the study of its feasibility. These approaches to ontology learning from text are generally based on the use of a corpus of texts. This corpus should be representative of the domain of the ontology. Using a set of techniques, we try to project in the ontology knowledge contained in texts by extracting concepts and relations.

We distinguish mainly five categories of text mining techniques:

- Linguistic techniques and lexico-syntactic patterns;
- Clustering techniques and / or classification techniques;
- Statistical techniques;
- Association rule based techniques
- and hybrid ones.

Besides of ontology learning from texts, ontology learning from Web appears to be a second category in domain-dependant Ontology.

The most known approaches exploit the textual Web content to enrich concepts using Word net Several approaches described in and enrich ontology's from Web documents.

Another approach is proposed in order to reduce the terminological and conceptual ambiguity among members of a virtual community. This approach proposes the discovery of concepts and relations from the Web sites and lead to the development by the system Onto Learn [13].

In these approaches, domain knowledge a priori is required. For this reason, they are dependent to the domain of the ontology and the collection of Web documents related to this domain need user intervention.

C. Incremental approach for Ontology learning from Web documents

On the other hand, other approaches are dedicated to the ontology building from Web, which is based on the generation of taxonomies without the use of knowledge or a priori or processing techniques of natural language and use of large corpus or thesaurus. The same approach were improved in [19] to an incremental approach of ontology learning from Web. In [19], a study of several types of available Web search engine and how they can be used to assist the learning process (searching web resources and compute IR measures). The learning process proposed by this approach is based on four steps:

- Taxonomic learning: the user starts to specify keyword used as a seed for the learning process from Web using a web search engine, the output of this step

is one-level taxonomy, a set of verbs appearing in the same context as extracted concepts.

- No-taxonomic learning: verb list and keywords are used as bootstrap for construction domain related patterns and to construct query to search engine.
- Recursive learning: The two previous learning stages are recursively executed for each discovered concept.
- Post-processing step consists in refining and evaluating the obtained ontology.
- This approach is domain independent and incremental. In the same context, our previous work was done. We have proposed an incremental approach of ontology learning from Web. We combined many text mining techniques and use an ontology-based IR System to classify the web documents.

D. Web structure mining-based approach for Ontology learning from Web

In [20], the underlying assumption behind this work is that the noun phrases appearing in the headings of a document as well as the document's hierarchical structure can be used to discover the concepts and taxonomic relations from documents.

A system that supports this approach is implemented and applied on a set of Arabic agricultural extension documents. It takes as input a root concept, analyzes all input documents' heading structure, extracts concepts from headings and builds a taxonomical ontology [21]

In this section, several approaches of ontology learning from web were detailed. Ontology extraction from texts belongs to this same work.

4. SOCIAL NETWORK ANALYSIS

Social network analysis deals with the interactions between individuals by considering them as nodes of a network (graph) whereas their relations are mapped as network edges. Social networks, such as Facebook and Bebo, are essentially online communities that allow users to come together, communicate and share things such as photographs, music or other files; and, most prolifically, to create short messages, often in the style of a mobile phone text message but shared among a group. People use the sites to ask their friends questions, say how they feel today and what they are up to, to comment on something they have seen on someone˝s page. A social network is the network of relationships and interactions among social entities such as individuals, groups of individuals, and organizations. Since the rise of Internet and the World Wide Web has enabled us to investigate large -scale social networks, there has been growing interest in social network analysis.

A social network is usually formed and constructed by daily and continuous communication between people and therefore includes different relationships, such as the positions, between ness and closeness among individuals or groups [22]. In order to understand the social structure, social relationships and social behaviors, social network analysis therefore is an essential and important technique. Research on social networks could be traced back to sociology, anthropology and epidemiology

Social Networks analysis and the direction of the research are therefore now moving from sociology to computer science. For social networks analysis, the analysis targets are mainly focused on resources from the web, such as its content, structures and the user behaviors. Application of data mining techniques to the World Wide Web, referred to as Web mining, can be used for the analysis of social networks [23]. In web mining, main analysis targets are from the World Wide Web, in the form of web content mining, web structure mining and web usage mining [24].

5. WEB TEXT MINING

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining [5], that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. CBIR will retrieve more similar image than the text retrieval. Content-based image retrieval (CBIR) is regarded as one of the most effective ways of accessing visual data [4]. To further improve the CBIR technique we have presented an improved algorithm by extracting feature vector comprises [8].

Image mining is an extension of data mining to image domain. Image mining is the concept used to extract implicit and useful data from images stored in the large data bases. Image mining is used in variety of fields like medical diagnosis, space research, remote sensing, agriculture, industries and even in the educational field [12].

The problem of Knowledge Discovery from Text (KDT) [6] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while

deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding.

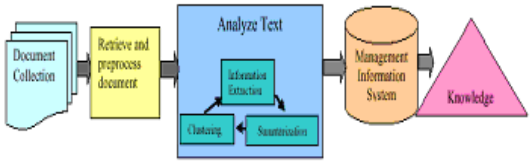

Figure-2. Text mining process.

Text mining applications

The main Text Mining applications are most often used in the following sectors:

- Publishing and media.
- Telecommunications, energy and other services industries.
- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Political institutions, political analysts, public administration and legal documents.
- Pharmaceutical and research companies and healthcare.

6. PREPROCESSING STEPS

In this paper, we can discuss the two crucial step of preprocessing namely Stemming and Stop word Removal. The overview of our system is depicted by the following figure.

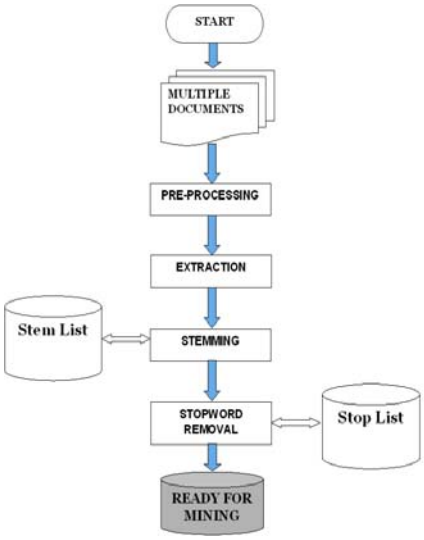

Figure-3. Preprocessing steps.

A. Extraction

This method is used to tokenize the filecontent into individual word.

B. Stemming

This method is used to find out the root/stem of a word for example, the words user, users, used, using all can be stemmed to the word "USE". The purpose of this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time. The stemming process is done using various algorithms. Most popularly used algorithm is "M.F. Porters Algorithm.

C. Stop word removal

Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no information. It may be of the following types such as pronouns, prepositions, conjunctions. Our system uses the SMART stop word list [4].

D. Effect of preprocessing

From the figure given below, it could be seen that the application of all the pre-processing techniques have a positive impact on the number of terms selected.

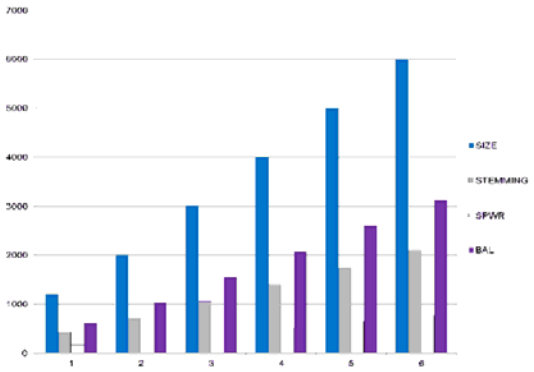

Figure-4. Effect of preprocessing.

7. TEXT CLUSTERING METHOD

Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies. Clustering of users will help to discover the group of users, who have similar navigation pattern. It's very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The clustering of page is useful for Internet search engines and Web service providers, since it can be used to discover the groups of pages having related content. The clustering operation is performed using social ontology. The social ontology has many relations and classes, the user conversion messages have top terms and class [2].

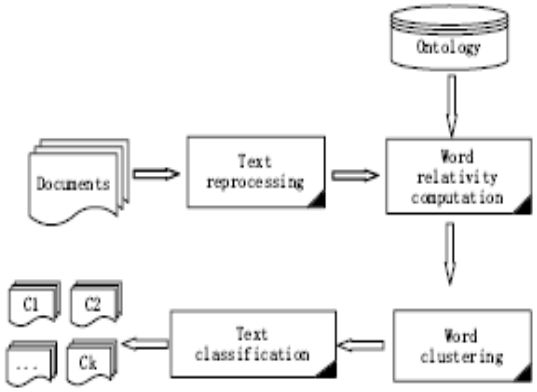

Figure-5. Clustering process.

The first step in text clustering is to transform documents, which typically are strings of characters into a suitable representation for the clustering task.

(i) Remove stop-words: The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). Remove stop-words can improve clustering results.

(ii) Stemming: By word stemming it means the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as work, worker, worked and working.

(iii) Filtering: Domain vocabulary *V* in ontology is used for filtering. By filtering, document is considered with related domain words (term). It can reduce the documents dimensions.

A central problem in statistical text clustering is the high dimensionality of the feature space. Standard clustering techniques cannot deal with such a large feature set, since processing is extremely costly in computational terms. We can represent documents with some domain vocabulary in order to solving the high dimensionality problem. In the beginning of word clustering, one word randomly is chosen to form initial cluster. The other words are added to this cluster or new cluster, until all words are belong to m clusters. This method allow one word belong to many clusters and accord with the fact. This method implements word clustering by calculating word relativity and then implements text classification.

Sequential pattern

This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. It's very meaningful for the Web marketer to predict the future trend, which help to place advertisements aimed at certain user group [3].

8. CONCLUSIONS AND FUTURE DIRECTIONS

The past five years have seen the emergence of Web Mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that the practicing it. In this paper we have briefly described the key computer science contributions made by the field, the prominent successful applications, and outlined some promising area of future research. One important focus is to enable search engines and other programs to better understand the content of Web pages and sites. This is reflected in the wealth of research efforts that model pages in terms of ontology of the content, the objects described in these pages

ACKNOWLEDGMENT

The authors would like to thank the Guide, reviewers and the editors for several suggested improvements.

REFERENCES

[1] LIU, B., "Web Data Mining: Exploring hyperlinks, contents, and usage data, Berlin Springer 2007.

[2] M. Mohamed Iqbal Mansur, Dr. C. Kavitha, Dr.K. Thangadurai, "Web User Profile Inference for User Group Interest Prediction on Social Networks using Domain Ontology", International Journal of Science and Research (IJSR), ISSN (online): 2319-7064, Impact Factor 3.358, Volume 3, Issue 8, August 2014, Page: 336-340.

[3] Dr. C.Kavitha, M. Mohamed Iqbal Mansur, Dr.K. Thangadurai, Survey on Web Mining Techniques and Applications", Journal of Computer Science and Applications, ISSN 2231-1270, Volume 6, November 1, 2014, Page: 411-416.

[4] Dr. K. Sakthivel, R. Nallusamy, Dr. C. Kavitha, "Image Retrieval Using Fused Features", World Academy of Science, Engineering and Technology International Journal of Computer, Information, Systems and Control Engineering. Vol. 8, No: 9, 2014.

[5] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers, 2006.

[6] McCallum, A., Corrada- Emmanuel, A., and Wang, X. 2005. "Topic and role discovery in social networks." In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. 786-791.

[7] Kimura, M., Saito, K., and Nakano, R. "Extracting Influential Nodes for Information Diffusion on a Social Network", ACM DMKD Explorations. vol. 20, issue 1, January, 2010, pp. 70-97.

[8] K. Sakthivel, T. Ravichandran and C. Kavitha, "Performance Enhancement in Image Retrieval Using Weighted Dynamic Region Matching", European Journal of Scientific Research. vol. 56, no. 3, pp. 385-395, 2011.

[9] Alfonseca E., Manandhar S. « An unsupervised method for general named entity recognition and automated concept discovery ». Actes de la 1er conférence internationale sur General Word Net. India, 2002.

[10] Allani H., Position paper: ontology construction from online ontologies International World Wide Web Conference archive Proceedings of the 15th international conference on World Wide Web.

[11] Aussenac-Gilles, Jacques M-P. Designing and Evaluating Patterns for Ontology Enrichment from Texts. EKAW 2006: 158-165).

[12] Sutha. S, Dr. C. Kavitha, "Automatic Image Annotation Using Semantic with Fuzzy KNN", Journal of Computer Applications (JCA), Volume VI, Issue 2, 2013.

[13] Missikoff, M., Navigli, R., and Velardi, P. (2002). Integrated approach to web ontology learning and engineering, IEEE Computer. Vol. 35(11) pp. 60-63.

[14] Faatz A. et Steinmetz R. Ontology enrichment with texts from the WWW. Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, 20th August 2002, Helsinki, Finland. 2002.

[15] Hahn U. et Markó K. Joint knowledge capture for grammars and ontologies. Proceedings of the First International Conference on Knowledge Capture K-CAP 2001, Victoria, BC, Canada, 2001.

[16] Hearst M.A. Automated Discovery of Word Net Relations. "Wordnet an Electronic Lexical Database". MIT Press, Cambridge, MA, 132-152, 1998.

[17] Hwang CH. Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information. Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), 14-20, 1999.

[18] Khan L. et Luo F. Ontology Construction for Information Selection. Proceedings of 14th IEEE International, 2002.

[19] Roux C., Proux D., Rechermann F. et Julliard L. An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. Proceedings of the ECAI2000 Workshop on Ontology Learning (OL2000), Berlin, Germany, 2000.

[20] Sekiuchi R., Aoki C., Kurematsu M. et Yamaguchi T., DODDLE: A Domain Ontology Rapid Development Environment, PRICAI98, 1998.

[21] Shiren Y. Tat-Seng C., Automatically Integrating Heterogeneous Ontologies from Structured Web Pages, Int'l Journal on Semantic Web and Information Systems. 3(2), 96-111, April-June 2007.

[22] Jin, Y. Z., Matsuo, Y. and Ishizuka, M. "Extracting Social Networks among Various Entities on the Web" In Proceedings of the Fourth European Semantic Web Conference, 2007.

[23] Chakrabarti, S. "Mining the Web: Discovering Knowledge from Hypertext Data" Morgan Kaufmann Publishers, USA, 2003.

[24] Cooley, R., Mobasher, B. and Srivastave, J. "Web Mining: Information and Pattern Discovery on the World Wide Web" In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, 1997, pp. 558-567, Newport Beach, CA, USA.