Ontology based Web Page Topic Identification

Abhishek Singh Rathore
Department of Computer Science & Engineering
Maulana Azad National Institute of Technology
Bhopal, India

Devshri Roy
Department of Computer Science & Engineering
Maulana Azad National Institute of Technology
Bhopal, India

ABSTRACT
With the emergence of the web, lots of research efforts are made in the area of Web Mining. This paper proposes an automatic approach for automatic topic identification from the web pages. The contribution of this research is in the approach of automatic topic identification of web pages that can provide better results. The topic of the web documents is identified through ontological approach. Keywords are extracted from the basic HTML tags and co-occurrence of words in the text instead of calculating the frequency of each term exits in a web page. Domain ontology is developed to map topics of the documents. Keywords are mapped to the ontology with a Levenshtein Edit Distance to extract topic of the web page. The result could give benefit to the search engines for faster tagging of web pages.

General Terms
Web Mining, Document Classification.

Keywords
DOM, Word co-occurrence, Ontology, Topic Identification.

1. INTRODUCTION
The World Wide Web contains billions of documents. These documents are from various sources like e-books, emails, news articles, research papers, and Web pages. The webpages may be small, may be large, written in different languages but covers all essential topics for Human.

With this huge amount of data, Web is usually unstructured or semi structured. Information extraction techniques are used to structure [1] this information. In order to structure the documents, it is necessary to understand the content of the document. Knowing the topic can help in understanding the content of the document. A topic [2], in an abstract sense can be thought as a subject matter, a category or simply a theme. Topic identification of a document assists search, background information gathering and contextualization tasks, and enhanced relevancy measures [3]. In simple terms, topic identification increases the precision of search.

A user may give short length query for searching documents. NEC Research Institute study [4] shows that up to 70% of searchers use only a single keyword i.e. the main theme as a search term. Single word queries given by users find matches to too many Web pages and hence the search results returned by the search engine include relevant as well as irrelevant pages. Categorizing web documents according to their topic can improve the precision of search results for single word queries.

The other major drawback of Web search is that the search results for a given query are independent of the domain in which the user made his request. Web document contains terms that may be belonged to different domains. For example, if a user wants to search a word "Charge", "Charge" may belong to domain of electricity or the domain of accountancy. Thus, the domain of the word should also be known for the topic identification to achieve higher relevancy.

This paper proposes an automatic topic identification algorithm based on domain ontology from the web documents. The process of topic identification includes keyword extraction from web documents. Keywords are extracted from the <title>, <meta> tags and headers of the web page (see Fig. 1). Although keywords present in HTML documents in Meta tag but it is not necessary that all web pages contain keywords in Meta tag and if they are present, there is no surety that they are correct. Hence along with keywords from HTML tags, keywords from the body content of the web page are also extracted for feature selection. Instead of most frequent words, most frequent co-occurring terms are extracted. Then the extracted keywords are mapped to the domain ontology to get the topic of the web document. Thus, the classification based on ontology has high accuracy, reliability and stability of information searching [5].

This paper is structured as follows: Section II explains the related works carried out in identification of topics. Section III describes the proposed automatic topic identification algorithm. Section IV provides results and analysis of the proposed work. Later sections provide the conclusion and future work.

2. RELATED WORK
Many authors had their proposed work in automatic topic identification of web documents. The majority of work based either on clustering of keywords/documents, or by machine learning or by ontological approach. In information retrieval, TF/IDF algorithm [6] is a very popular method to extract weighted index terms from the document. Keywords are extracted as follows:

Removal of Stop Words: In the text, connectives such as articles and prepositions that appear in a large number are known as stop words [7]. For example a, an, the, of, was, is etc. Since the occurrence of stop words is high but do not have any relevance to the documents, so stop words are removed from the text.

Word Stemming: Near about all natural languages contain parts of speech and tenses. For example, the words "go", "went", "gone", "goes", "going" have same meaning in text. In order to improve precision of search their root word "go" is to be identified. This process is known as stemming. Commonly used stemming algorithm is Porter Stemmer [7].

Assignment of numerical weights to each index term in the document vector: Different index terms have varying significance in describing the document contents. Some terms are more significant as compared to other terms. For example,

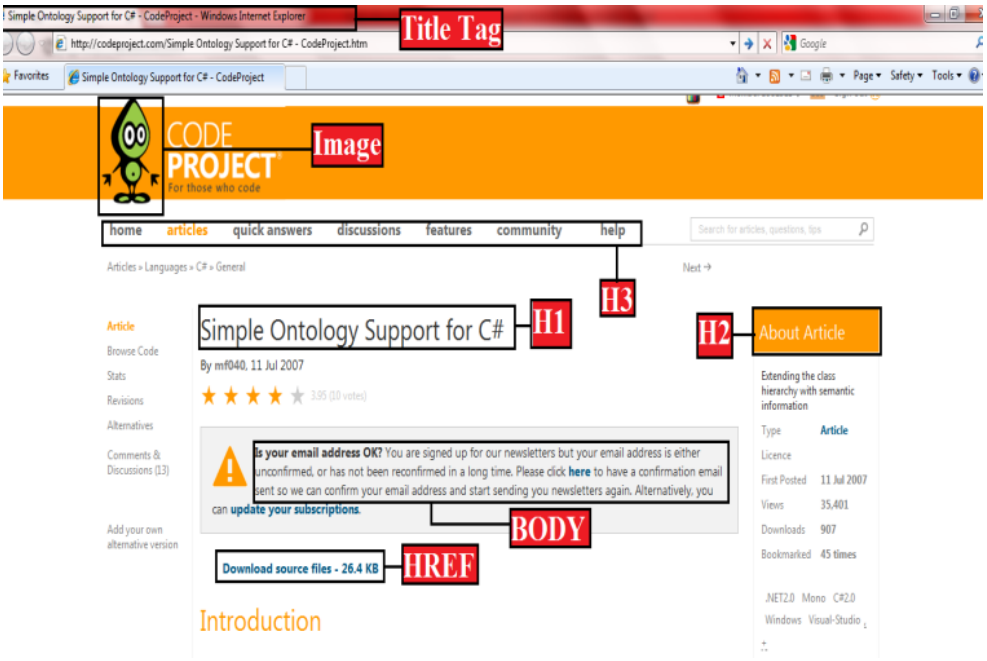

Fig 1: A sample Web Page with HTML tags

a document describing the topic "gravity", the term gravity is the most significant index term in that document. Therefore, each index term is assigned a numerical weight according to its significance in describing the document content.

However there is need to improve this algorithm and researchers are working for its improvement [8].

Closest to the proposed work is described by [9]. Keywords are extracted from Term Frequency formula and then keywords are mapped to domain ontology. Domain concept is identified through the ontology hierarchy. It provides better precision results than title tag based and TF value based approaches.

Another work exploits an ontological hierarchical structure [10] in order to find a topic of a text. The optimization process is carried out to shrink the ontology tree into an optimized tree to map keywords where only active concepts and the intermediate active concepts are chosen. The small size of optimized tree is reduced to a single path. This single path is retrieved using the Maximal Spanning Tree Algorithm.

Ontology-based web document classification and ranking method [11] is proposed for topic identification. Terms are extracted and ontology is built up, then similarity score between documents and ontology is computed based on WordNet by using Earth Mover's Distance (EMD) method.

A knowledge graph of encyclopedic concepts based on Wikipedia is built in [12], where the nodes in the graph are represented by the entities and categories. Link between encyclopedic graph and keywords, extracted from input text by Wikify ((Mihalcea and Csomai, 2007), is created.

Web page clustering is used for the automatic topic identification [13]. Clustering method is based on similarity metric which includes textual information, hyperlink structure and co-citation relations.

Baghdadi et al [14] describes that topics are identified in English text documents based on the chunks for each sentence in the document. Candidate topic is identified from the sentence using a noun phrase and the head of the verb phrase. The most weighted one candidate topic is selected as the main topic of the document.

Topics are identified by C. Wang et al [15] through Dirichlet Process Mixture Models. Gibbs sampling is applied to text in order to find a topic from the text.

Snasel et al [16] proposed that semi-discrete decomposition provides a smaller list of terms and these terms are mapped to the ontologies.

Burger et al [17] proposed classification of files based on ontologies. Classification process includes mapping of metadata of files with ontologies.

3. AUTOMATIC TOPIC IDENTIFICATION

This section describes the automatic topic identification system which consists of 2 parts: Keyword Extraction Subsystem and Ontology Mapping Subsystem (see Fig. 2).

3.1 Domain Ontology

Ontology formally represents knowledge as a set of concepts within a domain, and the relationships among those concepts [18]. It can be used to reason about the entities within that domain and may be used to describe the domain. There are lots of meaning arises from this definition. Firstly, ontology is domain specific; it does not contain knowledge of all areas but within a specific area. Secondly, ontology contains a number of classes/terms and hierarchical structure is used to define their relationships.

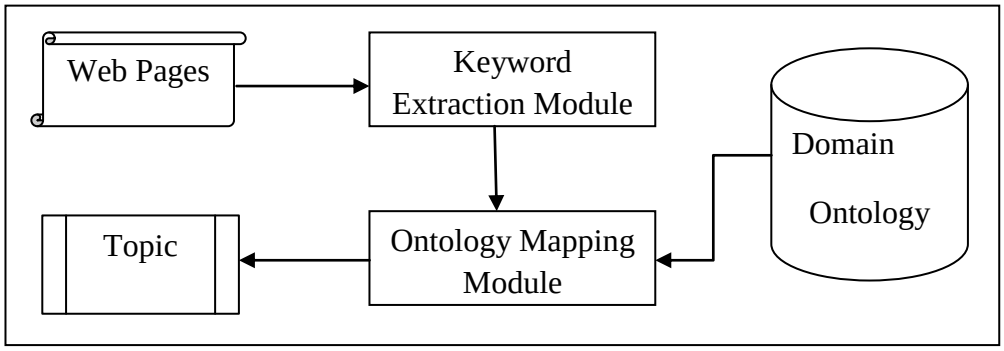

Fig 2: General Architecture of automatic topic identification system

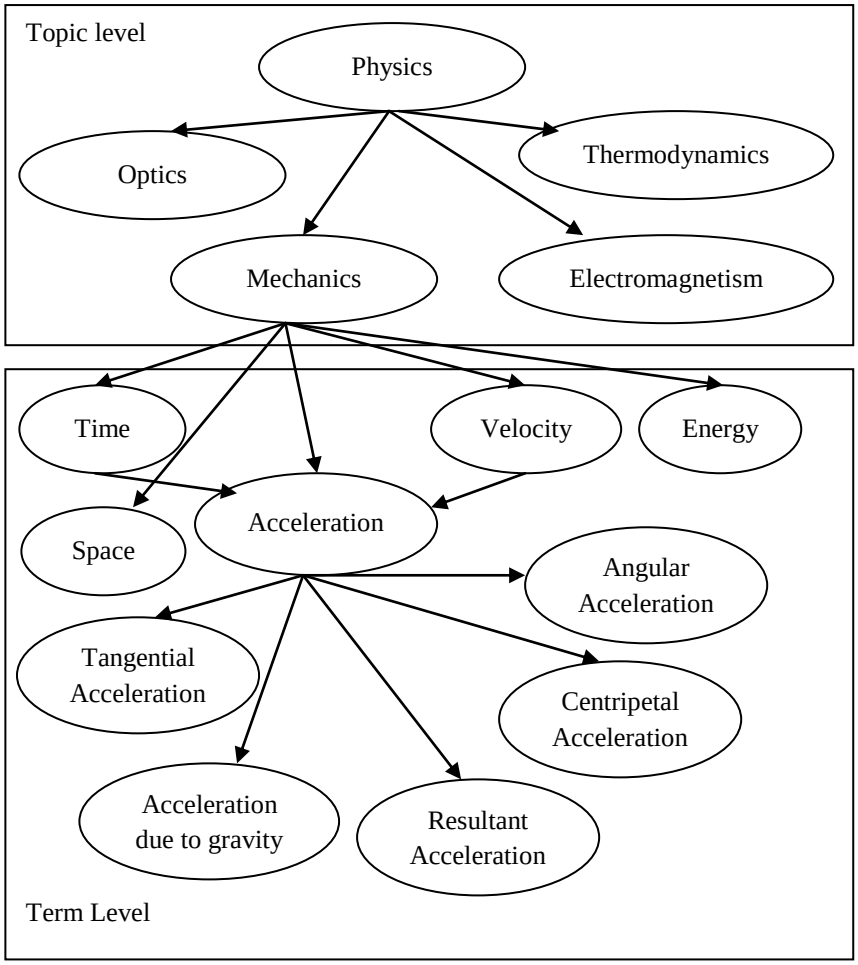

Fig 3: A part of the sample Ontology

Gruninger and Lee define the uses of ontology as follows (Gruninger & Lee, 2002,):

For communication between implemented computational systems, between humans and between humans and implemented computational systems

For computational inference: internally representing plans and manipulating plans and planning information as well as for analysing the internal structures, algorithms, inputs and outputs of implemented systems in theoretical and conceptual terms.

For reuse (and organization) of knowledge [19].

The domain ontology is developed manually with human knowledge base. The domain ontology is a two-level hierarchical structure. A part of the developed ontology used by the proposed system (see Fig. 3).

3.1.1 Topic Level

The topic level contains topics from a subject domain. A subject domain contains a number of topics. The topics may have many sub-topics. So, topics and subtopics share a parent child relationship. A subtopic may be placed under one or

more topics. The hierarchy of topics is stored as a tree structure with a single source.

3.1.2 Term Level

A topic consists of many terms. Terms form the next level of the ontology. A set of empirical relations is defined among the terms in a domain. The terms are the nodes of the graph. The edges between the nodes denote the relationship between terms. These relationships provide a means to infer possible semantic content of the textual documents. If a term is of significance in a document, it is usually the case that the document contains a number of references to related terms. In fact the occurrence of related terms is taken as a very strong indication of the relevance of the document.

The entities of the topic layer and the term layer are mapped according to relationships between entities of two layers.

3.1.3 Topic Term Layer Relationship

The documents on a topic contain several terms. A topic can explain more than one term. A term can belong to more than one topic. While keeping the topic layer to the term layer relationship, the terms covered by each of the topics of the topic taxonomy are kept separately.

3.2 Keyword Extraction

Table 1. Keyword Extraction Steps

S.No.	Steps
1	Parse the web page through the DOM and search the following HTML tags
1.1	<Title>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <meta> (description and keywords), <img>, <embed>
2	Extract the inner text of above mention tags
3	Remove the stop words
4	Let $H_{keywords}$ is a set of keywords extracted from HTML tags
5	Parse the tags and extract the inner text of tags
6	Remove the stop words
7	Extract keywords from text as described by [20]
8	Let F_k is the set of top K terms extracted by step 7
9	$Keyword_{union} = H_{keywords} \cup F_{keyword}$
10	$Keyword_{common} = H_{keywords} \cap F_{keyword}$

The problem of information extraction is to extract information from the contents of web pages. In information extraction, each web document is represented as a document vector. Document vector is a collection of keywords. In TF/IDF algorithm the number of occurrences of a term i.e. the term frequency is used for weight assignment. Only the term frequency is not sufficient for weight assignment, other factors like the consideration of the most significant terms within the HTML tag <keyword>, <meta> etc. are required to be considered for weight assignment. The HTML tag <title> contains the name given in that document. The name or title of the document is very important term and hence need to be considered for weight assignment. The actual texts of web pages are stored in various tags like <h1>, <h2>, <h3>, <h4>, <h5>, <h6>, <p>, <li>, <ul>, <tables>,
, <href > etc. These tags are parsed through Document object model (DOM). DOM represents a document in the form of a tree. Every node of the tree represents an HTML tag and hence inner-text within the HTML tag can easily be extracted through the tree structure. Each term of the inner text is stored as an index term in the document vector.

Keywords are selected as the method described by the [20] and from HTML tags (see Fig. 4). Frequent terms are extracted first, and then a set of co-occurrences between each term and the frequent terms, i.e., occurrences in the same sentences, is generated. Keywords are extracted as (see Table 1):

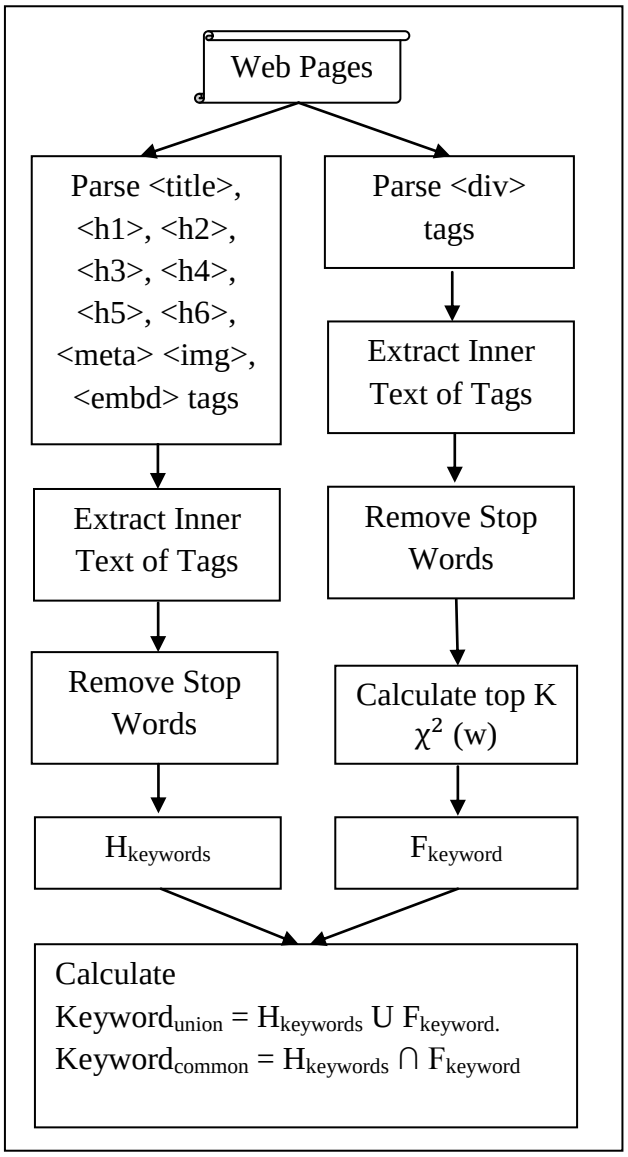

Fig 4: Keyword Extraction Module

3.3 Ontology Mapping

This section concerns with mapping keywords with ontology. Keywords extracted from the last section are mapped to the ontological concepts. There will be a possibility that all keywords of ontology will not mapped to the keywords extracted from web pages. Keywords may be different yet very similar. The Levenshtein distance algorithm has been used to measure similarity between keywords and match approximate strings with fuzzy logic.

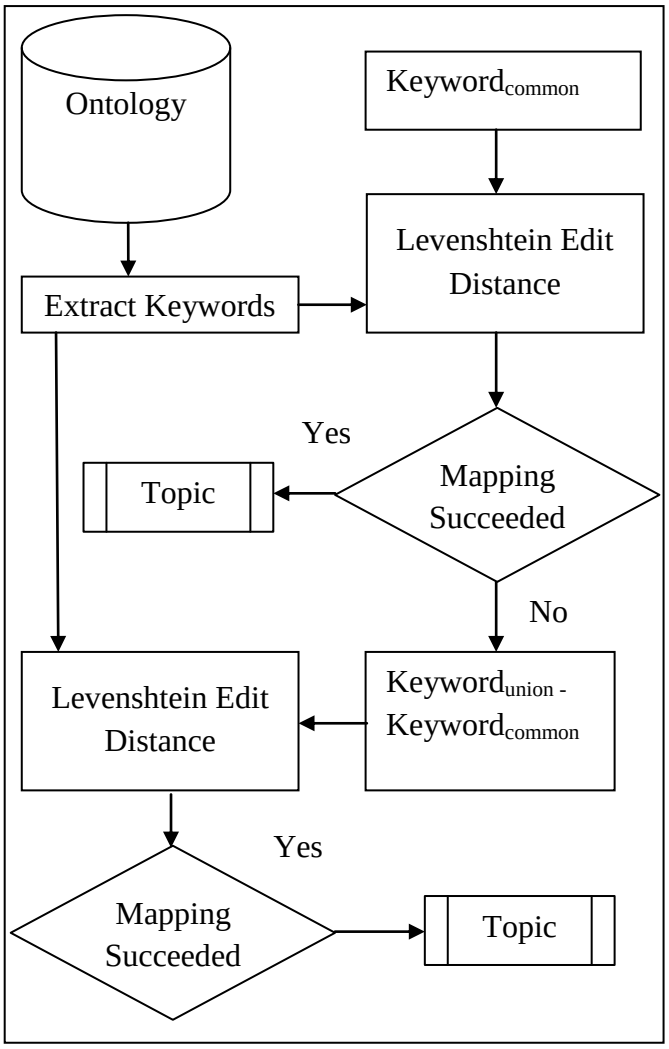

Fig 5: Ontology Mapping Module

In order to categorize topic, weights are assigned to the extracted keywords. Keywords from the set $Keyword_{common}$ have higher weights than the keywords from the set $Keyword_{union}$ - $Keyword_{common}$. The mapping procedure will be carried out in two phases. In the first phase, higher weight keywords are mapped. If the accumulated weight of the mapped keyword crosses threshold value than a topic is assigned to the web page, otherwise second phase is carried. In the second phase, again remaining keywords are mapped with ontology (see Fig. 5).

In order to map documents to the ontology, the ontology mapping module (see Fig. 5) works in two phases for two sets of keywords: $Keyword_{common}$ and $Keyword_{union}$. In the first phase, let K_{oi} is the set of keywords in the i^{th} node of ontology then $K_c = Keyword_{common}\ \sigma_0\ K_{oi}$, where σ_0 is Levenshtein Edit Distance of 0. If $|K_c| > 0.5|K_{oi}|$, it suggests that the document belongs to the i^{th} topic, otherwise, the ontology mapping module jumps to the second phase and looks for $Keyword_{union}$. If $|K_u| + 2|K_c| > |K_{oi}|$ then the topic is assigned to the document, where $K_u = Keyword_{union}\ \sigma_0\ K_{oi}$.

4. EXPERIMENTS AND RESULTS

This section describes the evaluation of the proposed approach for topic identification. For evaluation, experimental data contain documents from Wikipedia. Precision, Recall and F-measure formula is used to measure accuracy of the proposed approach.

$$Precision = \frac{hits}{hits + mistakes}$$

$$Recall = \frac{hits}{total\ documents}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

8924 documents of the Physics domain are crawled from the Wikipedia and keywords are extracted from these documents. Experiments are conducted for top 5, 10, 15 and 20 keywords and precision, recall and f-measure is calculated (see Fig 6).

A good average precision of 71.4% and recall of 40.5% is achieved in comparison of precision 69.85 by Tiun et. al [10]. The higher precision with lower recall suggests the improved performance of the proposed approach.

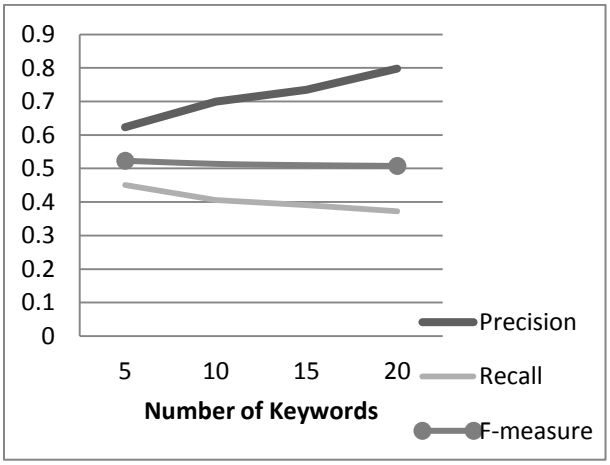

Fig 6: Precision, Recall, F-measure for different number of keywords

5. CONCLUSION

In this paper, the domain ontology is used for automatic topic identification of the web pages. Keywords are extracted from the HTML tags and the occurrence of the words in the text. The development of the domain ontology incurs costs in terms of both time and manual effort [21], but it gives higher precision.

It is a structure of topics and terms in a particular domain [21]. With higher precision and lower recall, the results show that the proposed methodology gives an efficient way of topic identification. With billions of web pages on the internet, it is not feasible to create ontologies for each topic and hence future work will include the topic identification without using domain ontologies.

6. REFERENCES

[1] Chang, C. H., Hsu, C.N. and Lui, S.C. 2003. Automatic information extraction from semi-structured Web pages by pattern discovery. Decision Support Systems. 35, 129-147.

[2] Villarreal, S. E. G., Elizalde, L. M. and Viveros, A. C. 2009. Clustering hyperlinks for topic extraction: an exploratory analysis. In Proceedings of the Eighth

Mexican International Conference on Artificial Intelligence.

[3] Coursey, K. and Mihalcea, R. 2009. Using Encyclopedic Knowledge for Automatic Topic Identification. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).

[4] Butler D. 2000. Souped-up search engines. Nature. 405, 112-115.

[5] Liu, X., Duan, X. and Zhang, H. 2012. Application of Ontology in Classification of Agricultural Information. In Proceedings of the IEEE Symposium on Robotics and Applications (ISRA).

[6] Salton, G. and Buckley, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management. 24 (5), 513-523.

[7] Chakrabarti, S. 2003. Mining the Web: Discovering the Knowledge from Hypertext Data. Elsevier Science, 48.

[8] Yang, Y., He, L. and Qiu, M. 2011. Exploration and Improvement in Keyword Extraction for News Based on TFIDF. In Proceedings of the ESEP 2011.

[9] Kong, H., Hwang, M., Hwang, G. Shim, J. and Kim, P. 2006. Topic Selection of Web Documents Using Specific Domain Ontology. In Proceedings of the MICAI 2006.

[10] Tiun, S., Abdullah, R. and Kong, T. E. 2001. Automatic Topic Identification Using Ontology Hierarchy. In Proceedings of the CICLing 2001.

[11] Fang, J., Guo, L., Wang, X. D. and Yang, N. 2007. Ontology-Based Automatic Classification and Ranking for Web Documents. In Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007).

[12] Coursey, K. Mihalcea, R. and Moen, W. 2009. Using Encyclopedic Knowledge for Automatic Topic Identification. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL).

[13] He, X., Ding, C. H. Q., Zha, H. and Simon, H. D. 2001. Automatic Topic Identification Using Webpage Clustering. In Proceedings of the International Conference on Data Mining (ICDM 2001).

[14] Shahsav, H. and Baghdadi, B. R. M. 2011. An Automatic Topic Identification Algorithm. Journal of Computer Science. 7 (9), 1363-1367.

[15] Wang, C., Yuan, C., Wang, X. and Xue, W. 2011. Dirichlet Process Mixture Models based topic identification for short text streams. In Proceedings of the Seventh International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2011).

[16] Snasel V., Moravec, P. and Pokorny, J. 2008. Using Semi-discrete Decomposition for Topic Identification. In Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications, ISDA '08.

[17] Burger S. and Stieger, B. 2010. Ontology-based classification of unstructured information. In Proceedings of the Fifth International Conference on Digital Information Management (ICDIM 2010).

[18] Available: http://en.wikipedia.org/wiki/Ontology_(information_science)

[19] Hepp, M. 2008. Ontologies: State of The Art, Business Potential, And Grand Challenges. In Ontology Management Semantic Web, Semantic Web Services, and Business Application, 7, Springer.

[20] Matsuo Y. and Ishizuka M. 2003. Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools. 10 (1), 157-169.

[21] Jain, S. and Pareek, J. 2010. Automatic Topic(s) Identification from Learning Material: An Ontological Approach. In Proceedings of the Second International Conference on Computer Engineering and Applications (ICCEA 2010).