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 knowledge, 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,
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.
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**

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parse the web page through the DOM and search the following HTML tags</td>
</tr>
<tr>
<td>1.1</td>
<td><code>&lt;title&gt;</code>, <code>&lt;h1&gt;</code>, <code>&lt;h2&gt;</code>, <code>&lt;h3&gt;</code>, <code>&lt;h4&gt;</code>, <code>&lt;h5&gt;</code>, <code>&lt;h6&gt;</code>, <code>&lt;meta&gt;</code> (description and keywords), <code>&lt;img&gt;</code>, <code>&lt;embed&gt;</code></td>
</tr>
<tr>
<td>2</td>
<td>Extract the inner text of above mention tags</td>
</tr>
<tr>
<td>3</td>
<td>Remove the stop words</td>
</tr>
<tr>
<td>4</td>
<td>Parse the <code>&lt;div&gt;</code> tags and extract the inner text of <code>&lt;div&gt;</code> tags</td>
</tr>
<tr>
<td>5</td>
<td>Remove the stop words</td>
</tr>
<tr>
<td>7</td>
<td>Extract keywords from text as described by [20]</td>
</tr>
<tr>
<td>8</td>
<td>Let $F_k$ is the set of top K terms extracted by step 7</td>
</tr>
<tr>
<td>9</td>
<td>$\text{Keyword}_\text{union} = \text{Hkeywords} \cup \text{Fkeyword}$</td>
</tr>
<tr>
<td>10</td>
<td>$\text{Keyword}_\text{common} = \text{Hkeywords} \cap \text{Fkeyword}$</td>
</tr>
</tbody>
</table>

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>`, `<table>`, `<br>`, `<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](image)

### 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.
In order to categorize topic, weights are assigned to the extracted keywords. Keywords from the set $\text{Keyword}_{\text{common}}$ have higher weights than the keywords from the set $\text{Keyword}_{\text{union}} - \text{Keyword}_{\text{common}}$. 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: $\text{Keyword}_{\text{common}}$ and $\text{Keyword}_{\text{union}}$. In the first phase, let $K_i$ is the set of keywords in the $i$th node of ontology then $K_i = \text{Keyword}_{\text{common}} \sigma_0 K_{\text{un}}$, where $\sigma_0$ is Levenshtein Edit Distance of 0. If $|K_i| > 0.5|K_{\text{un}}|$, it suggests that the document belongs to the $i$th topic, otherwise, the ontology mapping module jumps to the second phase and looks for $\text{Keyword}_{\text{union}}$. If $|K_i| + 2|K_{\text{un}}| > |K_{\text{un}}|$ then the topic is assigned to the document, where $K_i = \text{Keyword}_{\text{union}} \sigma_0 K_{\text{un}}$.

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.

\[
\text{Precision} = \frac{\text{hits}}{\text{hits + mistakes}}
\]

\[
\text{Recall} = \frac{\text{hits}}{\text{total documents}}
\]

\[
F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}
\]

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 Tiu et. al [10]. The higher precision with lower recall suggests the improved performance of the proposed approach.

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


Mexican International Conference on Artificial Intelligence.


