Semantic Document Classification using Lexical Chaining & Fuzzy Approach

Upasana Pandey, S. Chakraverty, Bhawna Juneja, Ashima Arora, Pratishtha Jain

Abstract: We propose a novel approach to classify documents into different categories using lexical chaining. In this paper we present a text categorization technique that extracts lexical features of words occurring in a document. Two kinds of lexical chains based on the WordNet and Wikipedia reference sources are created using the semantic neighborhood of tokens. The strength of each lexical chain is determined with the help of TF/IDF, category keyword strength and relative position of tokens in the document. Each category is assigned a weight depending upon the value obtained after the lexical chain computation. Fuzzy logic is incorporated to generate a range for each category using a triangular membership function. The document belongs to the category which satisfies the range criteria. Lexical chaining has large applicability in automated email spam filtering, topic spotting, email routing.

Keywords: Lexical Chaining, TF-IDF, Wikipedia, WordNet.

I. INTRODUCTION

The modern information age produces vast amounts of textual data, which can be considered largely as unstructured data. If it is properly organized and classified, then retrieving the relevant information from the maze of data becomes much simpler. With the exponential growth of documents, the need for automated methods to organize and classify the documents in a reliable manner becomes inevitable.

Text Categorization (TC), also known as Text Classification, is the task of automatically classifying a set of text documents into different categories from a predefined set [1]. Text-Classification (TC) [2] [3] methods can be classified as: Statistical and Semantic. Statistical methods consider words of a document as unordered or independent elements. These methods simply compute the frequency of the feature items. They do not take into account the characteristics of position and ignore the fact that words at different positions have different contributions to the theme of the article. The Term frequency–Inverse document frequency (TF-IDF) weight is often used in information retrieval and text mining. This weight is a statistical measure which evaluates how important a word is to a document in a collection or corpus [4]. Semantic methods [5] exploit the relationship among the words of a document in order to evaluate their semantic relevance.

Document classification refers to the process of sorting a set of documents into different categories. Organizations receive countless amounts of paper documents every day. These documents can be mail, invoices, faxes, or email. For companies to overcome the inefficiencies associated with paper and captured documents, they must implement an intelligent classification system to organize captured documents.

Traditionally, TC requires a substantial amount of manually labeled documents for classification which is often impractical in real-life settings. Keyword-based TC methods aim at a more practical setting. Each category is represented by a list of characteristic keywords which capture the category meaning. The effort is then reduced to providing an appropriate keyword list per category, a process that can be automated. Classification is then achieved by measuring similarity between the pre-defined category names and their keywords and the documents to be classified.

Cohesion and coherence are terms used in discourse analysis and text linguistic to describe the properties of written texts. Cohesion, “Connor writes”, is determined by lexically and grammatically over inter-sentential relationships, whereas coherence is based on semantic relationships.”Coherent text makes sense to the reader [6]. In cohesive and coherent text, successive sentences are likely to refer to concepts that were previously mentioned and to other concepts that are related to them. Lexical Cohesion can be defined as “the means by which texts are linguistically connected” (Carter 1998: 80). Often, lexical cohesion occurs not simply between pairs of words but over a succession of a number of nearby related words spanning a topical unit of the text.

Lexical Chaining is a technique which seeks to identify and exploit the semantic relatedness of words in a document. It is a process of identifying and grouping words together to form chains which in turn will help in identifying and representing the topic and content of the document [7]. The words of the text that make such references might be thought of as forming cohesive chains in the text. Each word in the chain is related to its predecessors by a particular cohesive relation. Lexical chains provide a clue for the determination of coherence and discourse structure, and hence the larger meaning of the text.

Manuscript Received October 25, 2011.

Upasana Pandey, Computer Engineering, Netaji Subhas Institute of Technology, Delhi University, New Delhi, India, 09871784621,
(e-mail: upasana1978@gmail.com).

S. Chakraverty, Computer Engineering, Netaji Subhas Institute of Technology, Delhi University, New Delhi, India, 09899568694,
(e-mail: apmahs@rediffmail.com).

Bhawna Juneja, Information Technology, Netaji Subhas Institute of Technology, Delhi University, New Delhi, India, 09910301135,
(e-mail: bhawnajuneja0705@gmail.com).

Ashima Arora, Information Technology, Netaji Subhas Institute of Technology, Delhi University, New Delhi, India, 09990495910,
(e-mail: ashima.arora91@gmail.com).

Pratishtha Jain, Information Technology, Netaji Subhas Institute of Technology, Delhi University, New Delhi, India, 09711214321,
(e-mail: peezenn@gmail.com).

We use lexical chaining to enhance the similarity percentage of belongingness. Our algorithm employs Wordnet and Wikipedia. The online lexical reference system called WordNet includes English nouns, verbs and adjectives organized into sets, each representing an underlying lexical concept. Similar meaning words are grouped together in WordNet in a synonym set called a synset. The different senses of a word represented in Wordnet are similarity through the use of synonyms, generalization of concepts through the use of hypernyms, specialized versions of a concept through the use of hyponyms or enunciation of parts of an object through meronyms. Lexical chains are formed between words belonging to the same synset. In WordNet, a word may belong to more than one synset, each corresponding to a different sense of the word. The relationship between two different words is determined by looking at all the senses of each of the words. Wikipedia is a free, open content online encyclopedia created through the collaborative effort of a community of users known as Wikipedians. It handles many fundamental tasks in computational linguistics, including word sense disambiguation, information retrieval, word and text clustering, and error correction.

The combination of Wordnet and Wikipedia provides us with a wide range of words which helps in forming lexical chains. The distribution and density of the chains in a text is an indication of coherence of the text. The strength of a lexical chain can then be used as a successful measure of the degree to which the similar-meaning words of the chain contribute to the overall meaning of the text.

II. BACKGROUND AND MOTIVATION

Information Science consists of having the knowledge and understanding on how to collect, classify, manipulate, store, retrieve and disseminate any type of information. Information science, in studying the collection, classification, manipulation, storage, retrieval and dissemination of information has origins in the common stock of human knowledge[8]. Information analysis has been carried out by scholars at least as early as the time of the Abyssinian Empire with the emergence of cultural depositories, what is today known as libraries and archives. The discipline of European Documentation emerged in the late part of the 19th Century together with several more scientific indexes. Paul Otlet, a founder of modern information science proceeded to build a structured document collection that involved standardized paper sheets and cards filed in custom-designed cabinets according to an ever-expanding ontology and a commercial information retrieval service. With the 1950's came increasing awareness of the potential of automatic devices for literature searching and information storage and retrieval. By the 1960s and 70s, there was a move from batch processing to online modes [9].

For text categorization using the context-oriented approach, a central concern is to automatically cull out a set of training documents from given corpus which can be associated with a certain category. Automatic generation of keyword-list per category is a potential area that motivates the use of contextual information to find suitable keywords for each category.

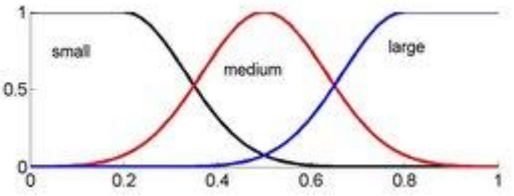
The Bag of Words (BoW) [10] scheme is a very popular scheme which has been used for representing documents. The BoW representation, using a simple frequency count alone, does not capture all the underlying information present in the documents. Moreover, it ignores information such as position, relations and co-occurrences among the words.

Lexical chains have been used as an intermediate representation of text for various tasks such as automatic text summarization [11], [12] malapropism detection and correction [13] and hypertext construction [14]. An algorithm for computing lexical chains was first given by [15] using the Roget's Thesaurus [16]. Since an electronic version of the Roget's Thesaurus was not available then, later algorithms were based on the WordNet lexical database.

In [17], the authors demonstrated how well organized background knowledge in form of simple ontologies can improve text classification results. Although designed primarily as a lexical database, WordNet can be used as ontology [18, 19, and 20]. For each concept present in a document, the referring terms can be found in WordNet starting from the relevant senses of the category name and transitively following relation types that correspond to lexical references. Thus, the concern about extracting a suitable list of keywords for a given category can be addressed by utilizing the WordNet.

The motivation of using Wikipedia is its ability to quantify semantic relatedness of texts. The basic purpose to include Wikipedia is to enhance the keyword list per category which is obtained from WordNet. This way terms which are practically related to the categories are also added. In addition, Wikipedia can be used as a machine learning tool [21].

The motivation of using fuzzy logic for text categorization comes from the fact that there is no clear separation between two or more categories and hence fuzzy logic using triangular membership function is a good way to deal with such fuzzy boundaries. A fuzzy set, A in a set of elements Z = {z} is characterised by a membership function, $\mu_A(z)$, that associates with each element of Z a real number in the interval [0,1].



Introduced in 1965 by Lotfi Zadeh, fuzzy logic has been used successfully in several areas like data mining and pattern recognition [22, 23, and 24]. Fuzzy logic [25] deals with fuzzy sets that allow partial membership in a set because a particular document may belong to more than one category.

In contrast to binary logic which defines crisp boundaries, fuzzy logic deals with the extent of relevance.

In this paper, we tap semantic information obtained from the WordNet [19] and Wikipedia [26][27] to ascribe contextual relationships between the words of a document to thus generate a set of keywords for each category. Lexical chains are formed on the basis of the words that are contextually related to each other i.e. fall in the same synset using WordNet, become part of the same lexical chain. Lexical weights are calculated and obtained results are combined with the weights, computed using semantic based approach. Finally combined results ascribe the document to its appropriate category using fuzzy logic. The concept of lexical chaining can also be used to separate junk mails from regular mails [28][29][30].

III. PROPOSED SCHEME FOR CLASSIFICATION

For the purpose of classification, we use a separate, set of training documents and a set of testing documents. For experimentation the corpus 20Newsgroups [31] is used; a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

The task of classifying the document is divided into two parts:-

A. *Keyword Extraction:* The first part exploits the semantic features encapsulated in the documents. We start with an initial set of categories. Category names are first input to the system. The WordNet lexical database is utilized to extract a set of keywords belonging to each category that possess strong semantic correlation with the category name. Wikipedia is used to make the keyword lists more exhaustive by adding first-level hyperlinks (related to the respective category names) to the category lists obtained from WordNet. Thus, each category is represented by a list of characteristic keywords which capture the category meaning and strengths associated with each keyword.

B. *Lexical Chaining:* The document is pre-processed to obtain tokens which signify the meaning of the document. There are two kinds of lexical chains based on the reference source are created using the semantic neighborhood of tokens; one obtained from WordNet and another from Wikipedia. Each chain is assigned a specific strength and contributes a certain amount towards the classification of the document to each category. A weight is assigned to each category depending on the relative belongingness of the chains to categories.

The *pseudo code* for the proposed algorithm is explained below:

1. *Acquiring keyword-list per category:* Words that are lexically related to a category (keywords) are collected from the WordNet source. First synonyms were extracted. Then hyperrnyms of each term are transitively located to allow generalization. Hypernyms are collected up to the first level only to avoid complexity. Next hyponyms at immediate lower level are located. Meronyms and coordinate terms are also considered to get an ontological list of each category. Then, Wikipedia is pinged with the category name. The list of first-level hyperlinks is created and the words which were absent initially were added to the category list.
2. *Pre-processing document:* The document to be classified needs to be pre-processed first. This includes:
   i. Position Storing: Positions of all the words in the document are stored before any processing.
   ii. Stop-word removal: All the stop words, i.e. words that appear frequently but do not affect the context are removed from the document. Examples of such words include ‘a’, ‘an’, ‘and’, ‘the’, ‘that’, ‘it’, ‘he’, ‘she’ etc.
   iii. Tokenizing: The document is fragmented into a set of tokens separated by some delimiters, e.g. whitespaces. These tokens (or terms) can represent words, phrases or any keyword patterns.
   iv. Stemming: The resulting set of tokens is replaced by their base form to avoid treating different forms of the same word as different attributes. This reduces the size of the attribute set. For example, both “celebration” and “celebrating” are converted to the same base form “celebrate”.
   v. Term Weighting: For each token w_i in the document’s token set, its frequency f_i is computed.
3. *Generating token-category relationship matrix:* We now construct the token-category relationship matrix. In this matrix number of rows is equal to the number of tokens in the document and number of columns is equal to total number of categories. The elements of this matrix denote the membership of a token to a category. The token membership $R_{i,k}$ of a token w_i to a given category c_k is obtained by $P_i(C_k)$; the presence (=1) or absence (=0) of a keyword w_i in the category C_k divided by the presence or absence of the same token in all the ‘n’ pre-defined categories.

   $$P_i(C_k) = \frac{\text{presence (1) or absence (0) of } w_i \text{ in } C_k}{\sum_{i=0}^{n} \text{presence (1) or absence (0) of } w_i \text{ in } C_i}$$

4. *Lexical Chaining:* Depending on the source, two kinds of lexical chains are created for the document- one obtained from WordNet references and another obtained from Wikipedia references. A token is added to a chain, only if intersection of its synset with the synset of a word in the chain is greater than a decided threshold i.e. they need to be semantically related words. Strength lc_i is calculated for each chain i depending on:

i. No. of words in the chain
ii. Cohesiveness(semantic distance between the lexical chain words)
iii. Relative strength of pairs of words(Tf-Idf)

This is formulated as,

$$CStrength = clength \times \left\{ \sum \frac{\sum ((Tf - Idf_{wi} \times S_{wi}) + (Tf - Idf_{wj} \times S_{wj}))}{distance\ between\ wi\ and\ wj} \right\} \div TotalNumberOfPairs$$

where CStrength is Chain Strength and clength is Chain length. Weights are assigned to each category based on the presence or absence of lexical chain words in its keyword list. At the end of this step, every category will have two different weights- one computed from WordNet reference and another computed from Wikipedia reference.

5. *Assigning document to a category:* The final weight of a category is the average of the two values. The document is classified to the category which satisfies the fuzzy triangular function.
6. *Wikipedia Training:* Training is used to enhance the ability of the system to correctly classify a document. A training document whose category is known is fed into the system. After pre-processing, a list of tokens is obtained. The tokens already occurring in the Keyword-Category list are sieved out and only the new ones are retained. The second-level hyperlink list of these tokens are found using Wikipedia. These lists are intersected with the respective category list. If the intersection results in a value greater than a specific threshold, those tokens are added to the Keyword-Category list resulting in an enriched category set.

The algorithm is summarized in Table I.

Table I: Pseudo code for Text Categorization

```
CTC( Categories C={$c_m$},Category Strength CSWordnet={$csWordnet_m$},Category Strength CSWiki={$csWiki_m$}, Category Strength $CS_m$={$cs_m$},Unlabelled Documents D={$d_j$}, Stopwords {$s_w$}, Tokens {T}, Semantics of tokens {S}, Features of document {F},Chain Strength from Wordnet LCWordnet={$lcWordnet_i$},Chain Strength from Wiki LCWiki={$lcWiki_i$}

Resources: WordNet,Wikipedia

Begin {
Step1: Acquire keyword list per category {
            2.3    {$c_m$}ϵ C,
               Acquire a representative keyword list using WordNet and Wikipedia.
    }
for each category $c_i$ do
{
 Generate triangular membership function values
   $meanc_i$ and $SDc_i$
   $meanc_i = ({}_{j=0}{}^{n}\sum cs_j)/n$
   $SDc_i = \{{}_{j=0}{}^{n}\sum(cs_j - meanc_i)^2\}/n$

}
Step2: Pre-process {
            2.1 Store the position of each token $w_{i.}$
            2.2  {dj} eliminate stopword $w_k$  in $d_j$ , if $w_k$  Î $d_j$  and {$s_w$} , then $d_j = d_j / w_k$
            2.3 Tokenize Document D
            2.4 Stemmer replaces inflected words with root forms.
     }
Step3: Generate token-category relationship matrix {
             3.1 Compute membership value $R_{i,k}$ as

/*where   $P_i$ ($C_k$)  is full (1) or no (0) membership of keyword $w_i$  in the  ontology list of category $C_k$  and  $N_k$ is number of keywords  */
    }
```

```
Step4:
       4.1 Lexical Chaining{
             for each token in the document do
               Identify lexical chains in global set1 with which the word has an identity/synonym/hyponym/ hypernym / meronym relation from WordNet lexical database.
                if No chain is identified then
                   Create a new chain for this word and insert in Global Set1.
                 end if
              Add word to the identified/created chains in the Global Set1.
          end for

for each token in the document do
               Identify lexical chains in global set2 with which the word has an identity/synonym/hyponym/ hypernym / meronym relation from Wikipedia reference source.
                if No chain is identified then
                   Create a new chain for this word and insert in Global Set2.
                 end if
              Add word to the identified/created chains in the Global Set2.
          end for

/*Global Set1 and Global Set2 signify the two types of lexical chains */

     }

      4.2 Chain Strength{
     for each category and words in Global Set1 do

$lcWordnet_i = clength \times \left\{ \sum \frac{\sum ((Tf - Idf_{wi} \times S_{wi}) + (Tf - Idf_{wj} \times S_{wj}))}{distance\ between\ wi\ and\ wj} \right\} \div TotalNumberOfPairs$

     for each category and words in Global Set2 do

$lcWiki_i = clength \times \left\{ \sum \frac{\sum ((Tf - Idf_{wi} \times S_{wi}) + (Tf - Idf_{wj} \times S_{wj}))}{distance\ between\ wi\ and\ wj} \right\} \div TotalNumberOfPairs$

           }

For each category i do
{
          $cs_i = (csWordnet_i + csWiki_i)/2$
 }

Step5: Assign document  D to category {
               D Î $c_m$  where $meanc_m - SDc_m \le cs_m \le meanc_m + SDc_m$
          }

Step6: Wikipedia Training {
       for each correctly classified document do
{
               For all tokens with  $P_i$ (C)  equal to zero,
               Second-level hyperlink lists are found using Wikipedia.
            These lists are intersected with the respective category list.
                  If the intersection is greater than a specified threshold, those keywords are added to the category list resulting in enriched category set.
}
```

IV. CONCLUSION AND FUTURE WORK

Our proposed scheme demonstrates that lexical chaining efficiently classifies a document into its most relevant category. For calculating the strength of each lexical chain we use not just the length of the lexical chain but other parameters such as relative position of words, Term Frequency-Inverse Document Frequency (Tf/Idf) and keyword strength as well. Chances of document misclassification are reduced for similar categories by using a triangular fuzzy membership function. This is an improvement over other methods which consider only the longest chain for text categorization and can be implemented in email spam filtering.

Even though the lexical chains manage to represent the semantics to a certain extent, although we feel, it can be further enhanced by more involved processing. Cataphora is used to describe an expression that co-refers with a later expression in the discourse of a text and Anaphora is an instance of a reference of preceding utterances. They could also be used as lexical features to achieve accurate document classification.

REFERENCES

1. Mohammed Abdul Wajeed, Dr. T.Adilakshmi, ”Text Classification using Machine learning”.A Journal of Theoretical and Applied Information Technology, Vol 7. No. 2, year 2009.
2. Yongguang Bao, Daisuke Asai1, Xiaoyong Du, Kazutaka Yamada1 and Naohiro Ishii1, “An Effective Rough Set-Based Method for Text Classification”, In Proc. Of IDEAL 2003, LNCS 2690, pp. 545-552, 2003.
3. Cheng Hua Li and Soon Choel Park ,“Text Categorization Based on Artificial Neural Networks”, In Proc of ICONIP 2006, Part III, LNCS 4234, p.p. 302-311, 2006
4. Hu Jin-zhu , Shu Jiang-bo,Huang Yu-ying , “Text Feature Extraction based on Extension of Topic Words and Fuzzy Set”.In Proc of 2008 Intl. Conference on Computer Science and Software Engineering.
5. Marco Ernandes et al, “ An Adaptive context Based algorithm For Term Weighting” , In Proc. of the 20th international joint Conference on Artificial Intelligence, 2748-2753 , 2007
6. http://www.cvisiontech.com/index.php?option=com_content&id=370&task=view&Itemid=312
7. Diwakar Padmaraju et al, “Applying Lexical Semantics to Improve Text Classification”, http://web2py.iiit.ac.in/publications/default/download/inproceedings.Pdf.9ecb6867-0fb0-48a5-8020-0310468d3275.pdf
8. http://en.wikipedia.org/wiki/Information_science.
9. Gerard Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Commun. ACM, 18(11):613–620.
10. Gregory Silber and Kathleen F. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics, 28(4):487–496.
11. Regina Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS’97), ACL, Madrid, Spain.
12. Graeme Hirst and David St-Onge. 1997. Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum, editor, WordNet:An electronic lexical database and some of its applications. The MIT Press, Cambrige, MA.
13. Stephen J Green. 1998. Automatically generating hypertext in newspaper articles by computing semanticrelatedness. In D.M.W. Powers, editor, NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language.
14. Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicatorof the structure of text. Computational Linguistics, 17(1):21–48.
15. Kirkpatrick. 1998. Roget’s Thesaurus of English Words and Phrases. Penguin.
16. Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press.
17. Stephen Bloehdron et al,“Boosting for Text Classification with Semantics Features”,In Proc of the MSW 2004 workshop at the 10th ACMSIGKDD Conference on Knowledge,Discovery and Data Mining, AUG (2004), p.p.70-87
18. http://www.webkb.org/interface/categSearch.html?categField=sports
19. http://www.WordNet-online.com/
20. Libby Barak et al, ”Text Categorization from Category Name via Lexical Reference” , In Proc. Of NAACL HLT 2009: Short papers, Pages 33-36, June 2009.
21. Building Semantic Kernels for Text Classification using Wikipedia by Pu Wang, Department of Computer Science, George Mason University,2007.
22. http://en.Wikipedia.org/wiki/fuzzylogic
23. Zadeh, L. A. 1973. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-
24. El-Sayed M. El-Alfy, Fares S. Al-Qunaieer, ”A Fuzzy Similarity approach for Automated Spam filtering”. In proc. of the 2008 IEEE/ACS International Conference on Computer Systems and Applications-Volume 00, Pages 544-550, 2008.
25. Choochart Haruechaiyasak, Mei-Ling Shyu, Shu-Ching Chen, “Web Document Classification Based on Fuzzy Association”, In Proc of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopmen, 2002.
26. www.wikipeida.org.
27. Learning to Link with Wikipedia by David Milne, Department of Computer Science, University of Waikato, 2008.
28. Q. Wang, yi Guan, X. Wang, “SVM Based Spam Filter with Active and Online Learning”, In Procs. of the TREC Conference, 2006.
29. I. Androutsopoulos et al., “Learning to filter spam email: a comparison of a naïve Bayes and a memory based approach,” In Procs of the workshop “Machine Learning and Textual Information Access”, 4th European Conference on Principles and Practice of Knowledge discovery in Databases, 2000.
30. Johan Hovold, “Naïve Bayes Spam Filtering Using Word Position Based Attributes”, International conference of Email and Anti spam, 2005.
31. http://people.csail.mit.edu/jrennie/20Newsgroups/