TopChat: encyclopedia-based topic identification from chat logs

Matías Nicoletti, Silvia Schiaffino and Daniela Godoy

ISISTAN Research Institute, UNCPBA - Tandil, Bs. As., Argentina
CONICET, National Council for Scientific and Technical Research - Argentina
{mnicolet,sschia,dgodoy}@exa.unicen.edu.ar

Abstract Textual conversations on the Internet, such us chat rooms or instant messaging services, have become an excellent source of data for semantic analysis. In particular, potential user interests or user-related topics could be extracted from these conversations for personalization purposes. In this work, we present a novel method for topic detection from chat logs. First, we defined the generic structure of the process. Then, a variety of text-mining techniques was evaluated in each step of the process. Stemming, synonyms, POS tagging and named entities recognition are examples of these techniques. Encouraging experimental results from a comparative evaluation procedure, allow us to determine the most suitable combination of techniques for the problem.
Keywords: *topic identification, semantic analysis, text mining, chat, encyclopedia knowledge, concept association.*

1 Introduction

The search for user interests has become a matter of concern in the personalization of information contents. In general, people tend to talk about the topics in which they are interested in. In fact, one of the possible sources for extracting implicit information is conversations: the current development of social tools for digital interaction provides a rich source of data for semantic analysis. Through an automatic text processing system, relevant topics could be mined from textual conversation in order to create user profiles with interest information.

Natural Language Processing is known as a prominent research area in Computer Sciences. In particular, the semantic analysis through topic identification has acquired significant relevance in the last years. Even though there is a variety of works, topic identification is still an immature research area. Most related works aim to identify topics on general purpose documents; our study targets noisy text sources, such us chat room logs. A variety of strategies have been applied for detecting topics from text, like frequent term vectors [1], named entities recognition [2] and concept association [3,12,10]. For inferring the semantic meaning of word and phrases, different sources of knowledge have been proposed: Open Directory Project [1], Wordnet[2,12], Wikipedia [3,6,4,7], among others. Also, there have been approaches that use classification-based methods for the

task [9]. However, using a classifier to assign topics to documents is not a scalable solution when the knowledge source is extended to the size of Wikipedia.

In this context, we propose a novel approach called *TopChat* (TOPic identification from CHAT logs). It consists of an unsupervised method for detection of topics from chat data logs. The main idea is to associate Wikipedia articles, considered as concepts of human knowledge, to text messages in order to infer the topics of conversation. Since chat logs are a source of data characterized for being noisy, we strongly emphasized the text pre processing stage. Therefore, we considered different combinations of techniques for data cleansing. Then, two strategies were proposed for the association of concepts to user messages: (i) using the raw text of messages for a search in the dictionary and (ii) identifying entities from messages previous to the search for concepts. Through empirical evaluation, the different combinations of pre processing techniques with concept association strategies were compared. For the experiment, we defined a metric called *relevance score*, related to the relevance of the set of concepts linked to user messages. In this way, we were able to determine the best combination with promising results for a novel approach in this area of study.

The rest of the article is organized as follows. Section 2 presents a detailed description of the proposed approach, considering the dictionary, and the two different strategies for tackling the topic identification problem. Section 3 describes the evaluation procedure and discusses the results obtained. Section 4 summarizes the related works found in the literature. Finally, in Section 5 we present our conclusions and future work.

2 Proposed approach

In this work, we introduce *TopChat*, an unsupervised method for topic identification. The general schema is presented in Figure 1. The system inputs are mainly text logs from chat conversations. In the first step of the process, the input is parsed in order to identify the involved users and their messages. As a result, messages are grouped by users and irrelevant content from the logs (e.g. log timestamps) is filtered out. Secondly, noisy text is pre processed. As it is expected, the text of chat logs is extremely dirty. Therefore, data cleansing techniques are used to prepare the user messages for future analysis.

The third step is the concept association. For this purpose, we use a semantic dictionary containing concepts of human knowledge. The result of this step is a set of concepts associated to each message. This information is expected to give a general idea of the semantic meaning of the messages.

Finally, a category hierarchy is built from each concept. The information about categories is also extracted from the dictionary. First, we connect each concept with its corresponding first level category. Next, the process is repeated for higher level categories. As a consequence, we are able to identify topics from text considering different levels of abstraction.

The final result of the process is a set of user profiles, each one containing (1) a ranking of the most relevant concepts, (2) a ranking of the most relevant

first level categories and (3) a list of the most significant general categories of any level. Since this profile is based on topics that are regularly mentioned by the user, the information may be considered as the user interests or, at least, user-related concepts.

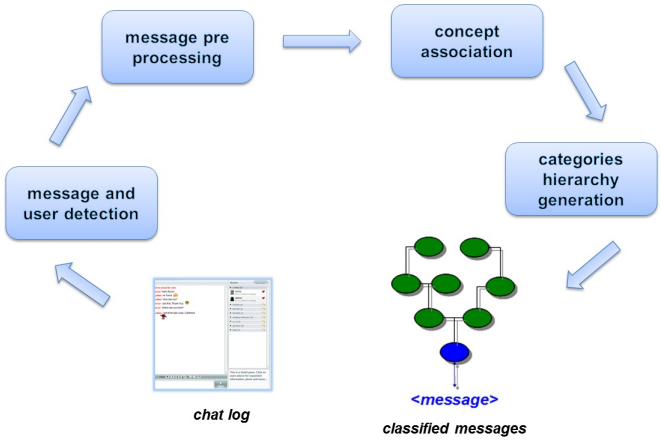

Figure 1. Overview of our proposal

The rest of this section is organized as follows. Section 2.1 describes the dictionary used to perform the semantic analysis. Then, we propose two approaches based on the general schema in order to tackle the topic identification problem, which are described in Sections 2.2 and 2.3.

2.1 Wikipedia as the dictionary

One of the main sources of knowledge in the Web is Wikipedia[1], which consists of a large set of articles (over 3M), multi-level categories and disambiguation information that describes a variety of concepts of human knowledge. Since Wikipedia contains an extremely high amount of unstructured information, performing any kind of analysis becomes a hard task. Therefore, we took advantage of DBpedia, which is a community project aiming at the extraction of structured information from Wikipedia. From DBpedia website[2], users are able to download a complete snapshot of the encyclopedia in several parts, depending on their information needs. Taking advantage of this remarkable feature, we used the information about article titles, extended abstracts, categories relations and term disambiguation.

[1] See http://www.wikipedia.org/
[2] See http://dbpedia.org

For each article, we indexed with Lucene[3] the title and extended abstract in order to generate a concept index. Similarly, we built two indexes for categories: one index relates articles with first level categories and the other one relates first level categories with higher level categories. Also, a disambiguation index was developed to handle ambiguous concepts.

When using Lucene, text analyzers are commonly used tools for the pre processing of text before indexing and searching. To evaluate the most suitable techniques, we defined 4 kinds of analyzers: (1) the standard *analyzer*, which executes the most common techniques like *stop-words* filtering, lower case conversion and URL detection, (2) the stemming *analyzer*, which adds Porter's stemming algorithm[4] to the *standard* process, (3) the *synonym analyzer*, which adds synonyms for each term using Wordnet to the *standard* analyzer, (4) the *synonym-stemming analyzer*, which first adds synonyms and then, applies stemming to the *standard* process.

2.2 First approach: using the raw text of messages

Based on the general schema, we proposed a first approach for topic identification. The general structure of our method (Section 2) must be specified with concrete operations and techniques in each step. No changes are made to the first step, since the input is parsed in order to identify the users and messages. For the second step, we defined the pre processing strategy with the aim of handling the chat text issues. This involves (i) the deletion of references to users by their names, (ii) the filtering of invalid characters and (iii) the execution of analyzer operations, like *stop-words* filtering or the stemming algorithm.

As regards the third step, we defined a specific strategy for concept association. In this case, we used the concept index built from Wikipedia articles. The strategy is simple: the whole text of each message is used to perform a search similar to the one done in search engines (based on the typical Term Frequency x Inverse Document Frequency measure) in the index. The first N items of the result sets are matched to each message. In addition, the result position is considered for establishing a relevance value the concept.

Finally, the two category indexes are used for the categories hierarchy generation. Initially, we associated first level categories to concepts, and then, we established relations with the higher level categories in order to build the hierarchical structure. Since categories tend to become too general (e.g. LIVING PEOPLE) and the computing time grows exponentially, we decided to limit the hierarchy tree depth to 3 levels. An example of this process can be found in Figure 2.

2.3 Second approach: identifying entities from messages

After conducting informal tests on the first approach we realized that some possible improvements could be made. Therefore, we proposed a second ap-

[3] Apache Lucene website. http://lucene.apache.org/

[4] See http://snowball.tartarus.org

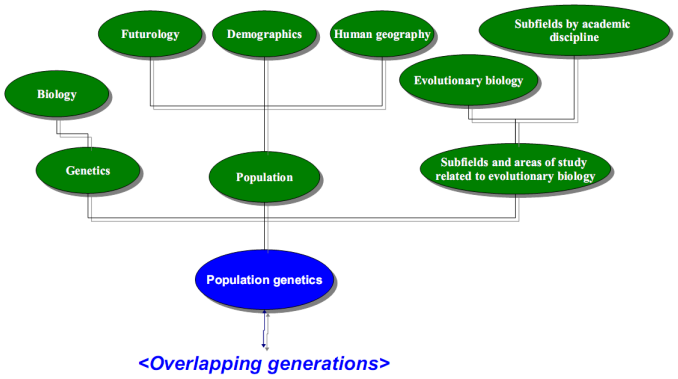

Figure 2. A sample category hierarchy for the concept *Overlapping generations*

proach based on the detection of relevant entities from the messages. The process presents some differences with the first one (Figure 3). Also, we used two typical text mining tools: a named entities recognizer and a POS (Part-Of-Speech) tagger, both provided by the Standford NLP Group[5].

A named entities recognizer typically uses a classification-based approach to detect named entities, like persons, institutions, locations or any kind of proper nouns. A POS tagger is a tool to automatically assign a grammatical label to every word in a sentence. In particular, we are interested in the analysis of nouns and their modifiers, like adjectives or adverbs (adverbs are actually adjective modifiers). By grouping the results of both techniques, a set of entities is associated with each message. In contrast to the first approach, the entities names are used as the search parameter in the concepts index. Consequently, we have a higher number of concepts in each profile.

We also identified some inefficiencies in the dictionary. Since a significant number of articles in Wikipedia are related to ART, in particular to MOVIES and MUSIC, some of the titles are similar to common expressions used by people in conversations. *That is a good idea* or *it is ok* are examples of this situation, which we considered *irrelevant messages*. We noticed that art topics are frequently linked with a user profile, although the user is not referring to them. Therefore, an optimization of the indexes was performed in order to filter this kind of information , which is not relevant to our domain of study.

Additionally, we realized that most irrelevant messages do not provide important information for topic identification. Thus, the pre processing step was modified in order to add a filtering process of messages containing less than 4 words. Although this is a simple approach that must be improved with a more sophisticated method, it is useful to handle the irrelevant messages issue.

[5] See http://nlp.stanford.edu/

Eventually, ambiguous concepts are associated to entities. In these cases, DBpedia provides a useful disambiguation data base extracted from Wikipedia articles. We used an adapted version of Micheal Lesk's algorithm [8], which considers a windows of the N nearest concepts to the ambiguous one (considering N=2). Thus, the description of each disambiguation concept is compared to the nearest concepts descriptions using the cosine text comparison function. Finally, the most related concept replaces the ambiguous one.

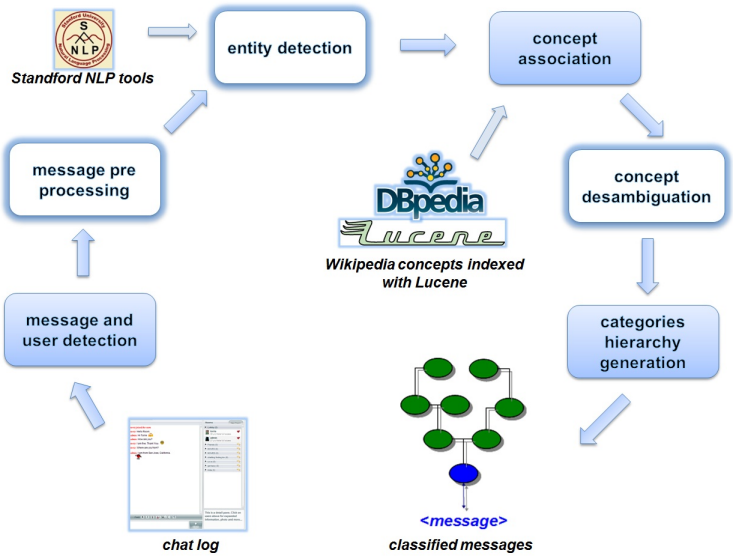

Figure 3. Overview of the entity-based approach (modified components in white)

3 Experimental evaluation

The empirical evaluation of the proposed approaches was divided in 2 stages. In the first stage, we designed a test procedure to determine the most suitable Lucene analyzer according to the domain characteristics. To start with the test, we selected a chat log about overseas resources in Australia from the Society of Genealogist[6] as the sample input data. The log contains 323 messages from 17 different users. Secondly, we executed on the input an algorithm implementing the first approach. In fact, we ran the algorithm 4 times, using in each execution one of the analyzers defined in Section 2.1. As a result, we obtained 4 sets of user profiles: each set produced with one different analyzer. We identified the user who made more contributions to the conversation (52 messages). Next,

[6] See http://www.sog.org.uk/prc/australasia.shtml

we manually assigned one relevance score to each message, which represents the relevance level of the 3 associated concepts. The score can have 3 possible values: 0, if none of the 3 concepts is relevant to the message; 0.5, if at least one concept is moderately relevant to the message; 1, if at least one concept is completely related to the message. Also, for each message we recorded the position of the most related concept, which is a measure commonly known as *hit position*. It must be noticed that when the relevance score is 0, the hit position is not defined.

Since the main goal is to evaluate and compare the effectiveness of each variant, we defined 4 metrics based on the data collected from the test. First, the *relevance score TOP 1* is the average of the relevance scores for each message, but just considering the first associated concept. Similarly, *relevance score TOP 2* is the average of the relevance scores, but considering the 2 first concepts. In the *relevance score TOP 3,* the 3 associated concepts are considered. Finally, we defined *AVG hit position*, which is the average of the hit positions for each message. The results of the test case execution for each analyzer are summarized in Figure 4 and Figure 5.

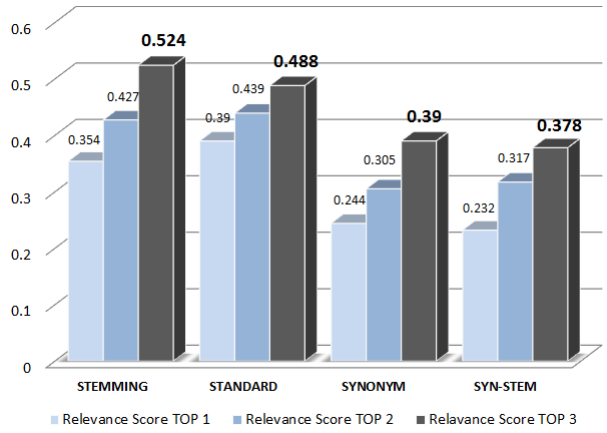

Figure 4. Comparison of the relevance scores (first-approach-based method)

The results show that both *synonym* and *syn-stem* analyzers have a considerable lower relevance scores than the others. Since both analyzers use a synonym-based technique, we may assume that the use of synonyms increases the total number of concepts that are associated, but reduces the quality of the first ones. Because the objective of the system is to associate at least one relevant concept, these analyzers are not suitable for topic identification. On the other hand, *standard* analyzer has the highest *TOP 1* and *TOP 2* relevance scores, as well as its *AVG hit position* is the closest to 1. Therefore, relevant concepts are frequently detected in the first 2 positions. However, the analyzer with the high-

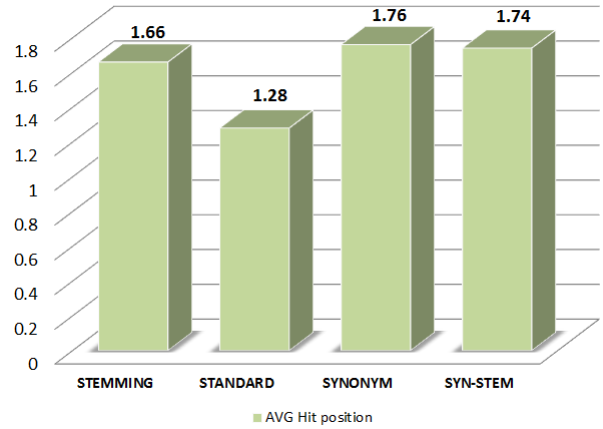

Figure 5. Comparison of the AVG hit positions (first-approach-based method)

est *relevance score TOP 3* value is *stemming*, which exhibits the best general effectiveness.

The second stage of the evaluation has the objective of comparing the first and the second approach. Therefore, we designed a second test that uses an implementation of the entity-based method. The procedure is similar to the first evaluation, as it has the same general structure, the same input and uses the same metrics. The main difference is that both relevance score and hit position were not assigned to each message. Instead, we manually assigned one relevance score and one hit position for each entity detected in each message, called *entity relevance score* and *entity hit position*. Then, we computed the *message relevance score* as the average of the entity relevance scores for each message, and the *message hit position* is the average of the entity hit positions. Finally, *relevance score TOP X* metrics were calculated in the same way as the first test, but using the *message relevance scores*. Similarly, the *AVG hit position* is the average of the *message hit positions*. In Figure 6 and Figure 7 we present the results only for the *standard* and *stemming* analyzers, which have previously shown the best results. In this occasion, we considered the same user with 52 messages and 101 entities.

The second experiment shows that, although the *AVG hit position* is slightly lower, the *stemming* analyzer has higher relevance scores than *standard*. Additionally, the entity-based method combined with the *stemming* analyzer achieves the highest *relevance score TOP 3* of the whole evaluation procedure, with 0.65. Finally, we conclude that the second approach has a higher *relevance score TOP 3* than the first approach, with an increase of 0.13 for *stemming* and 0.08 for *standard*.

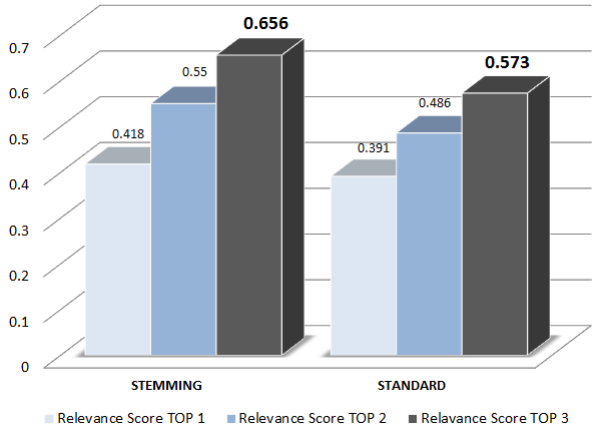

Figure 6. Comparison of relevance scores (second-approach-based method)

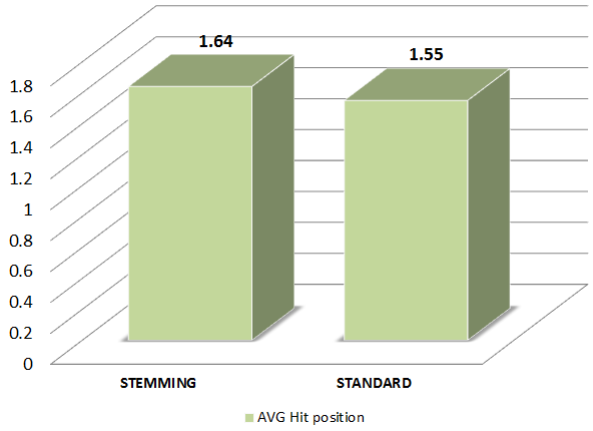

Figure 7. Comparison of AVG hit positions (second-approach-based method)

4 Related work

We have detected certain common aspects among most of the works related to our research. As regards the general methodology, we found (i) approaches based on frequent term vectors, commonly known as *bag-of-words* (BOW) [1], (ii) approaches that only use the detection of named entities [2], (iii) techniques using concepts rather than just terms [3,12,10]. With respect to the general knowledge source, Wikipedia has become the most popular alternative [3,6,4,7]. This rich encyclopedia has been used as a concepts and/or categories dictionary for semantic analysis. Other alternatives used for this purpose are the manually created Yahoo! directory[7][12], Open Directory Project[8] [1] or Wordnet[9] (considering the hypernyms and hyponyms relations) [2,12].

To the best of our knowledge, there are not previous researches on topic identification in noisy texts (e.g. chat logs) using Wikipedia as the general knowledge source. The most similar work is *Wikify!* [4], in which Wikipedia concepts are associated to sets of documents using a keyword detection algorithm. Nevertheless, some differences are identified: (i) the keyword extraction algorithm is based on the ranking of possible n-grams extracted from the dictionary, while our similar entity-based approach uses a combination of tools like a POS tagger and a named entities detector; (ii) just the Wikipedia article titles are considered, whereas our method also considers the extended abstracts; (iii) it is a general purpose document approach, while our approach aims for informal text from chat.

Another closely related study is presented in [3]. Initially, *Wikify!* is used to detect concepts and then, a Wikipedia-graph centrality algorithm is applied to discover related topics. An advantage of this method is that not mentioned topics could be detected. However, this strategy reflects negatively on the accuracy as not relevant concepts may be associated, depending on the quality of the links between the topics in the graph. In [11] an approach with the same objective as ours is introduced. In this case, the whole text of the input documents is matched with Wikipedia articles texts using the cosine similarity function. A similar situation is reported in the study in [10], where the approach presented uses the complete text of the input document and only considers the Wikipedia articles titles in the dictionary.

In contrast with our unsupervised approach, the technique presented in [9] requires previously annotated data to infer the topics from an input document. Similarly, in [5] the authors introduce a text classification system that calculates the most relevant Wikipedia concepts to a given input document. The classification approach works fine when only a limited number of topic and/or categories are considered. However, serious difficulties will be met if all the possible topics in Wikipedia are used. Also, in [2] clustering is performed over detected named entities in order to find representative topics in documents. In this case, Word-

[7] See http://dir.yahoo.com/
[8] See http://www.dmoz.org/
[9] See http://wordnet.princeton.edu/

net is used to generate a small set of keywords in order to improve the entity recognition. Nevertheless, this approach is only tested on a limited knowledge source that is considerable smaller than Wikipedia.

5 Conclusions and future work

In this paper we presented a novel technique for automatic topic identification from noisy text. Initially, the system aims for text logs from chat rooms, but it could be easily extended to instant message conversations. Our approach takes advantage of some of the text mining technologies, like POS tagging and named entities recognition, as well as it exploits the semantic power of the current hugest knowledge source in the Web, Wikipedia. A comparative evaluation of potential text analyzers was carried out, concluding that the best analyzer was *stemming*. We concluded that the use of strategies like synonyms, increase the number of associated concepts, which is related to the recall metric used in Information Retrieval. Indeed, recall is not an important factor to be targeted in this study. Additionally, two different approaches for topic identification were proposed. We empirically established that the entity-based alternative is superior to the one using the complete texts of messages. In fact, the method combining the *stemming* analyzer with the entity-based approach has a general relevance score of 0.65. This means that in the 65% of the times, the system associates relevant concepts to messages from an extensive dictionary of about 3M possibilities.

Since encouraging results were obtained, we propose future work to continue with this research. In order to improve the general effectiveness of the algorithm, we could evaluate the use of techniques for handling (i) writing mistakes (specially mistakes that represent valid words), (ii) not relevant messages (replacing the current strategy for a more sophisticated one), and (iii) abbreviations (for instance, considering if an abbreviation refers to a term already used in the context). Also, more efforts must be invested in the optimization of the dictionaries. In particular, the dictionary contents could be tailored to the specific analysis needs, instead of using a general purpose concept data base. A better precision is expected to be achieved if the irrelevant concepts are eliminated from the dictionary. For instance, if the analysis domain refers to software development, then just the software engineering concepts and categories must remain in the indexes.

Acknowledgments

This work has been partially supported by ANPCyT (Argentina) through PICT 2007 Project Nº 529 and by CONICET through PIP Project Nº 114-200901-00381.

References

1. J. Bengel, S. Gauch, E. Mittur, and R. Vijayaraghavan. Chattrack: Chat room topic detection using classification. In H. Chen, R. Moore, D. D. Zeng, and J. Leavitt,

editors, *Intelligence and Security Informatics*, volume 3073 of *Lecture Notes in Computer Science*, pages 266–277. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25952-720.

2. C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. *IEEE Trans. on Knowl. and Data Eng.*, 16:949–964, August 2004.
3. K. Coursey, R. Mihalcea, and W. Moen. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 210–218, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
4. A. Csomai and R. Mihalcea. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23:34–41, September 2008.
5. E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1301–1306. AAAI Press, 2006.
6. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
7. T. Kliegr. Entity classification by bag of wikipedia articles. In *Proceedings of the 3rd workshop on Ph.D. students in information and knowledge management*, PIKM '10, pages 67–74, New York, NY, USA, 2010. ACM.
8. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM.
9. O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia, 2008.
10. P. Schonhofen. Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.
11. Z. S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, editors, *ICWSM*. The AAAI Press, 2008.
12. S. Tiun, R. Abdullah, and T. E. Kong. Automatic topic identification using ontology hierarchy. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '01, pages 444–453, London, UK, 2001. Springer-Verlag.