Expert Systems

Article

DOI: 10.1111/exsy.12133

Conceptual feature generation for textual information using a conceptual network constructed from Wikipedia

Amir H. Jadidinejad,[1] Fariborz Mahmoudi[1] and M. R. Meybodi[2]

(1) Faculty of, Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
E-mail: amir.jadidi@qiau.ac.ir
(2) Computer Engineering and Information Technology Department, Amirkabir University of Technology, Tehran, Iran

Abstract: *A proper semantic representation of textual information underlies many natural language processing tasks. In this paper, a novel semantic annotator is presented to generate conceptual features for text documents. A comprehensive conceptual network is automatically constructed with the aid of Wikipedia that has been represented as a Markov chain. Furthermore, semantic annotator gets a fragment of natural language text and initiates a random walk to generate conceptual features that represent topical semantic of the input text. The generated conceptual features are applicable to many natural language processing tasks where the input is textual information and the output is a decision based on its context. Consequently, the effectiveness of the generated features is evaluated in the task of document clustering and classification. Empirical results demonstrate that representing text using conceptual features and considering the relations between concepts can significantly improve not only the bag of words representation but also other state-of-the-art approaches.*

Keywords: conceptual feature generation, bag of concepts, document classification, document clustering, Wikipedia mining, semantic space, random walks

1. Introduction

The analysis of information retrieval systems using standard benchmarks such as Text REtrieval Conference[1] has shown that in spite of growing complexity, no system is considerably superior to the others (Armstrong *et al.*, 2009). Following the similar line of research, the same results have been reported for document classification (Gabrilovich & Markovitch, 2009).

Most of the corresponding techniques represent the text as bag of words (BOW), which only accounts for term frequency in the documents and ignores important semantic relationships between key terms that limit its usefulness to represent the ambiguities in natural language (Ferragina, 2011; Nasir *et al.*, 2013). Co-occurrence analysis (Sahlgren, 2006; Turney & Pantel, 2010) (i.e. latent semantic analysis (LSA) (Landauer & Dumais, 1997)) has been proposed to identify the existence of relations between key terms. However, these models can only be reasonably used when texts are fairly long and perform sub-optimally on short texts.

A good deal of recent work is attempting to go beyond this paradigm by enriching the input text with conceptual annotations (Ferragina, 2011). In these models, a semantic annotator is responsible to enrich the representation by leveraging the structure of a background conceptual network (Jadidinejad & Mahmoudi, 2014; Jadidinejad *et al.*, 2015). Noteworthy, these models are different in the nature of the concept, which are ranged from explicit human-organized (Gurevych & Wolf, 2010; Jadidinejad *et al.*, 2015) to collaboratively generated concepts (Gabrilovich & Markovitch, 2009; Gurevych & Zesch, 2013; Hovy *et al.*, 2013).

Concept-based models represent textual documents using semantic concepts, instead of (or in addition to) keywords, and perform analysis in that concept space (Egozi *et al.*, 2011). It is clear that representing textual documents using high-level concepts will result in a representation model that is capable of handling both synonymy and polysemy.

Therefore, the contributions of this paper are twofold. First, an algorithm is proposed to generate conceptual features for textual information with the aid of a prior knowledge source; using these knowledge-based features, we can detect the topical semantics of the input text, and the proximity of meaning between texts is derived from the notion of proximity between the concepts corresponding to those texts. Furthermore, various tools have been proposed for automatic conceptual network construction and exploration. Second, the applicability of the proposed model is demonstrated on two fundamental natural language processing (NLP) tasks, document clustering and classification.

The remainder of this paper is organized as follows. In Section 2, related works are reviewed comprehensively. Section 3 describes the methodology and different parts of the proposed system. Section 4 explains how the proposed

[1] http://trec.nist.gov/

approach is employed upon for document clustering and classification. Finally, conclusion and future works are presented in Section 5.

2. Related work

The typical representation model for document clustering and classification is the BOW paradigm. Its main restriction is that it assumes independency between words and ignores all the conceptual relations between them. Concept-based models have attempted to tackle this problem by using statistical or knowledge-based approaches.

Statistical approaches build a ‘semantic space’ [also known as (aka) ‘word space’ or ‘distributional semantics’] of words from the way in which these words are distributed in a corpus of unannotated natural language text (Sahlgren, 2006; Turney & Pantel, 2010). There are many types of semantic space algorithms that go beyond this simple co-occurrence model (Jurgens & Stevens, 2010).

The most successful and well-known technique in this field is LSA (Landauer & Dumais, 1997), which relies on the tendency for related words to appear in similar contexts. Although latent concepts can represent the topical semantic of the input text better than visible words, they might be difficult to interpret in natural language. Moreover, polysemy is not captured because each occurrence of a word is treated as having the same meaning due to the word being represented as a single point in the space without considering its context.

Despite the statistical features, a number of knowledge-based lexical/conceptual features have been proposed in the literature. The underlying repositories were mostly WordNet or Wikipedia (Gurevych & Wolf, 2010). Gabrilovich and Markovitch (2009) proposed explicit semantic analysis that leverages concepts explicitly defined by humans. Explicit Semantic Analysis (ESA) represents textual information with respect to the external–textual article space [such as Wikipedia (Gabrilovich & Markovitch, 2009), Wiktionary, Open Directory Project (ODP) (Gabrilovich & Markovitch, 2007) or any comprehensive collection of textual articles (Anderka & Stein, 2009)]. In addition, it is also able to indicate how strongly a given word in the input text is associated with a specific article in the external space. Based on this model, two pieces of texts can be semantically related in spite of having no word in common. It is believed that the most important aspect of ESA is its strength in utilizing an external–textual collection (article space) to generate explicit features for a fragment of text. Different experiments (Anderka & Stein, 2009) have shown that the nature of the external article space (ESA concept hypothesis (Agichtein *et al.*, 2009)) is not significant at all.

Explicit or latent features generated by ESA or LSA have the potential to address synonymy and *somehow* polysemy, which are arguably the two most important problems in NLP. On the other hand, these approaches are inherently limited in some aspects. E/LSA generates a *static* feature vector for a *word* irrespective of its context. In other words, they disregard word order, and therefore disambiguation will be a serious problem in such cases. To suppress the effect of this problem, combining different feature vectors has been proposed (Gabrilovich & Markovitch, 2009). This problem will be more serious when dealing with long text documents. The experimental results have shown that long documents and ambiguous words are poorly represented in these models. Egozi *et al.* (2011) applied ESA to weak queries of text retrieval conference collection and found that the generated concepts are too noisy.

Using external knowledge sources (specially Wikipedia) for enriching classic document representation was introduced by Gabrilovich and Markovitch (2009). In the following line of research, Wang and Domeniconi (2008) and Hu *et al.* (2008, 2009c) applied conceptual features on the task of classification and clustering, respectively. Recently, Huang *et al.* (2012) used 17 different features to train a document similarity measure on Wikipedia and WordNet. In this model, each document is represented at both the lexical and semantic levels.

Additionally, Yazdani and Popescu-Belis (2013) proposed a document similarity measure using the visiting probability from one set of nodes to another. Both content and structure of Wikipedia have been leveraged to generate visiting probabilities. The proposed similarity measure has been evaluated by different important tasks in NLP and information retrieval.

3. Generating conceptual features for textual information

We address the problem of enhancing the classical BOW representation by enriching the input text using concepts of an external concept repository. Unlike previous researches (Gabrilovich & Markovitch, 2009; Fodeh *et al.*, 2011; Szyma & Ski, 2014) that implicitly extending the classic BOW representation model with additional features corresponding to concepts, we identify in the input text short-and-meaningful sequences of terms (aka spots (Ferragina, 2011)) that are then linked to unambiguous entities drawn from an external concept repository. The most interesting benefit of the proposed model is the structured knowledge attached to the input document that leverages not only a bag of concepts but also the valuable structure defined between concepts.

In this paper, Wikipedia link graph is employed as a conceptual network, and a semantic annotator is used to detect topical semantic of the input text. Figure 1 illustrates the logical flow of the proposed model. At the first phase, the Wikipedia link structure is mined to build the background conceptual network (see Section 3.1 for more details). The second phase contains two components: ‘wikifier’ (Milne & Witten, 2013) that extracts explicit seed concepts occurred in the input text and ‘semantic annotator’ is responsible to enrich the representation by leveraging the structure of the conceptual network. The semantic

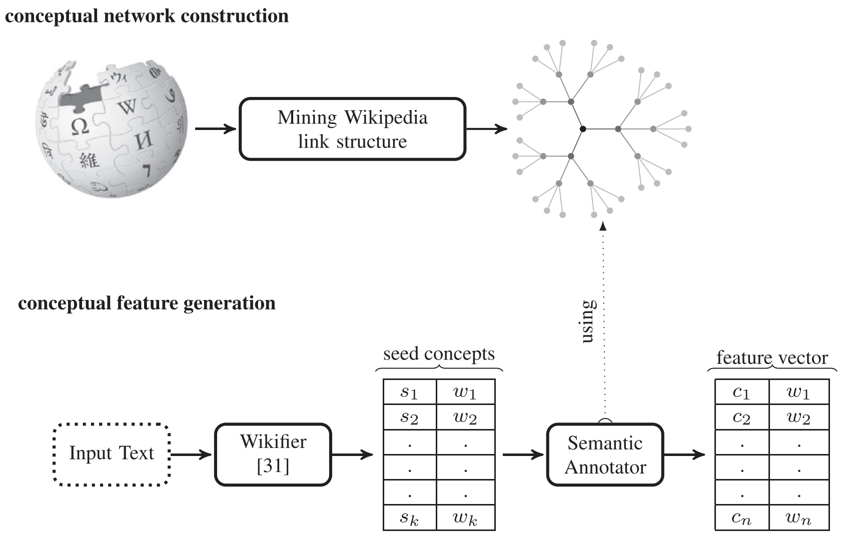


Figure 1: *Logical flow of the proposed model.*

annotator represents a piece of text as a conceptual feature vector,[2] which contains new generated features that represent the topical semantic of the input text better than the ones supplied with the explicit words (Jurgens & Stevens, 2010; Turney & Pantel, 2010) or the implicit/explicit concepts (Landauer & Dumais, 1997; Agichtein *et al.*, 2009).

In the following subsections, the specific knowledge acquisition method is described in Section 3.1. It exploits Wikipedia to build a comprehensive conceptual network. Moreover, the corresponding semantic annotator for generating conceptual features for the input text is described extensively in Section 3.2.

3.1. Conceptual network construction

Regardless of Wikipedia's obstacles such as textual form and human-level applicability (Gabrilovich & Markovitch, 2009), a growing community of researchers has recognized it as a valuable repository of concepts (Medelyan *et al.*, 2009). As a matter of fact, different parts of Wikipedia such as internal links, category structure, textual contents, disambiguation pages and revision histories (Medelyan *et al.*, 2009) have been leveraged in various researches. Following successful experiments (Hu *et al.*, 2009a; Yeh *et al.*, 2009; Navigli & Ponzetto, 2012; Nastase & Strube, 2013), all articles, categories and internal links between them were utilized to produce a set of concepts and their corresponding associated links.

Notably, the existence of a link between two articles does not always imply a semantic relation. Despite the writing rules in Wikipedia (Medelyan *et al.*, 2009), some phrases are just linked because there are entries for the corresponding concept. Therefore, to assure topic relatedness of association links, it is assumed that two articles are associated if they are linked to each other in Wikipedia. Moreover, bidirectional links have been leveraged in a previous research (Hu *et al.*, 2009a). Other kinds of internal links such as links between articles and parent categories ('art-cat') and hierarchical links between categories ('cat–cat) have also been leveraged without revision.[3] Therefore, each article or category is assumed as a concept. Articles with equivalent category have also been merged together. Figure 2 presents the structure of article and category graph and the way they are combined together. Noteworthy, category graph can be used to produce generalized concepts, while article graph provides more specific ones.

Although using the larger amount of knowledge leads to small improvements in NLP tasks (Zesch & Gurevych, 2010), Wikipedia snapshot of 20 November 2007 was used in order to have comparable experiments with previous researches (Gabrilovich & Markovitch, 2009). After parsing the Wikipedia XML dump using WikipediaMiner (Milne & Witten, 2013), about 2 million articles and categories are selected as concepts. Each concept has 7.67 associations in average. The most common concept is 'American film actors' with 9117 neighbours.

3.2. Semantic annotator

The role of semantic annotator is to extend the input seed concepts using the structure of conceptual network. This remark represents the task of semantic annotator as a classic network theory problem: 'finding nodes that are semantically related to a given node'. This is a well-studied problem, and two different approaches have been proposed in the literature. Some researchers (Hughes & Ramage, 2007;

[2] A weighted sequence of Wikipedia concepts that is sorted by the relevance to the context of the input text.

[3] See (Medelyan *et al.*, 2009) for more information about different kinds of links in Wikipedia.

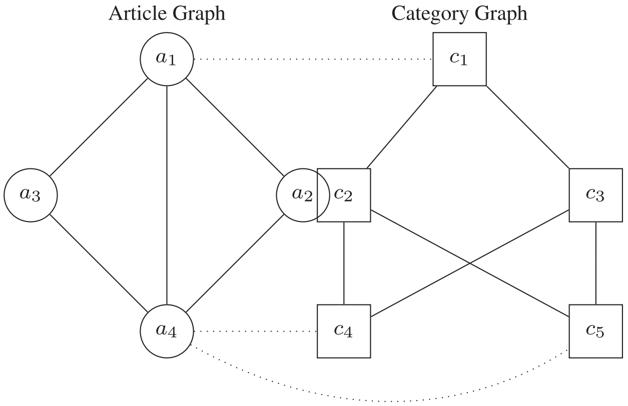


Figure 2: *Combining article graph and category graph to build the background conceptual network. Each solid edge corresponds to a bidirectional link between two articles or categories, while each dotted edge corresponds to a parent category of a specific article.*

Ollivier & Senellart, 2007; Hu *et al.*, 2009a; Ramage *et al.*, 2009; Yeh *et al.*, 2009) represented the underlying network as a Markov chain and tried to employ stochastic techniques to predict a conceptual neighbourhood around node *i*. On the other hand, other researchers (Syed *et al.*, 2008; Gouws *et al.*, 2010) leveraged a heuristic search using spreading activation algorithms on the network.

Spreading activation algorithms (Gouws *et al.*, 2010) is initiated by labelling a set of source nodes with *activation*, which is a numerical value intended to represent some aspects of a composite concept or textual information. Over time, when a node becomes active, the activation is propagated to other nodes via the association links between them like a chain of falling dominoes. Each round of such a propagation is called a *pulse*. In contrast with stochastic solutions, although heuristic methods require less computational resources, they are highly sensitive to the initial distribution and different parameters. In this section, following previous successful researches (Hughes & Ramage, 2007; Ramage *et al.*, 2009; Yeh *et al.*, 2009; Zhang *et al.*, 2010), the focus is on stochastic approaches, and a hybrid solution is proposed by customizing personalized PageRank (Haveliwala, 2003) with the ideas coming from spreading activation and Green measure (Ollivier & Senellart, 2007).

After constructing a comprehensive network of the concepts and relations between them, the first step of semantic processing is mapping the text to a set of concepts appearing in the text. 'Entity linking' is the task of linking entities mentioned in text with their corresponding entities in a conceptual network (Wang & Han, 2014). The emergence of knowledge-based approaches in different applications of information retrieval caused entity linking to receive much more attention in recent years (Wang & Han, 2014). WikipediaMiner's wikifier (Milne & Witten, 2013) is one of the most accurate and accessible tools in this field (Wang & Han, 2014). In this paper, wikifier was leveraged to map the input document into a weighted set of seed concepts.

In the following subsections, the problem of semantic annotation is formalized in Section 3.2.1; then, the proposed algorithm and its corresponding parameters approximation are presented in Sections 3.2.2 and 3.2.3, respectively.

3.2.1. Formalization Let p_{ij} be the transition probabilities of a Markov chain on a finite set of states *C*, which is a non-negative number representing the probability to jump from node $i \in C$ to node $j \in C$. These transition probabilities form a stochastic matrix, the so-called transition matrix (*T*). Because our background conceptual network is an undirected graph, the simple random walk can be defined on it by setting $p_{ij} = 0$ if there is no edge from *i* to *j*, and $p_{ij} = 1/d_i$ if there is an edge from *i* to *j*, where d_i is the degree of *i*. Personalized PageRank (aka topic-sensitive PageRank) (Haveliwala, 2003) biases the global PageRank method by focusing the PageRank computation around some seed subset of nodes. The output of personalized PageRank is an *a priori* estimation of vertex importance regarding the seed concept:

$$P_{t+1} = (1-\lambda)T{\cdot}P_t + \lambda S \qquad (1)$$

At each step, it is possible either to follow an association link with probability $1-\lambda$ or jump back to a seed concept (*S*) with probability λ (teleport operation). This algorithm takes the activation distribution from the current time step (P_t) and multiplies it to the transition matrix (*T*) in order to produce a better estimation for the next time step (P_{t+1}). For any ergodic Markov chain, there is a steady-state probability vector that is the principal left eigenvector of *P* (Haveliwala, 2003). Previous researches (Ollivier & Senellart, 2007) have shown that the probabilities generated by personalized PageRank are overly influenced by the underlying graph connectedness and tend to reproduce the global values generated by the original PageRank algorithm. It means that the general concepts always get more weight irrespective of the selected seed concepts. For example, 'American film actors' (a concept with the highest degree in the concepts network constructed in Section 3.1) obtained considerable weight when personalized PageRank algorithm was applied by starting with 'every' set of the initial seed concepts independently of the fact that the initial set of seed concepts relates to 'technology' or 'politics'. Noteworthy, Green measure (Ollivier & Senellart, 2007) is a way to get rid of this problem:

$$P_{t+1} = T{\cdot}P_t + S - v \qquad (2)$$

The Green method attempts to break the tendency of personalized PageRank to general concepts by penalizing each concept with its query-independent equilibrium measure (*v*) (Ollivier & Senellart, 2007), which represents the global importance of each concept in the link structure of the graph, regardless of the selected seed concepts. When the system has reached equilibrium, the highest values in the estimation vector (P) refer to those concepts that are the most related ones to the source concepts and have a global importance in the link structure.

According to this algorithm, the nodes' weights are reduced in each iteration based on their centrality to solve the problem of the convergence of personalized PageRank algorithm to concepts with high centrality. Therefore, the weights of the nodes with very high centrality are reduced considerably in every iteration of the algorithm, and other nodes will have the opportunity to emerge. On the other hand, related high centrality concepts are charged in each iteration through more nodes, and this decay factor will have lower effect on them.

In the original Green algorithm (Ollivier & Senellart, 2007), PageRank value of each node has been used to determine global centrality of the nodes in the network [v in equation (2)]. According to this issue, whereas determining of PageRank values for all nodes available in the network requires hundreds of iterations of random walks over the entire network, high complexity of employing Green algorithm will be its prominent problem. The previous researches in the field of knowledge-based information retrieval (Hughes & Ramage, 2007; Hu *et al.*, 2009a; Ramage *et al.*, 2009; Yeh *et al.*, 2009; Zhang *et al.*, 2010; Yazdani & Popescu-Belis, 2013) have almost used personalized PageRank algorithm to determine the related concepts and have not considered the problem of the convergence of the algorithm to general concepts. In the following subsection, the proposed algorithm for ranking nodes in a network according to a specific initial set of seed nodes is presented. The authors believe that paying attention to the problem of general concepts convergence in random walk algorithms (especially personalized PageRank) is one of the reasons for better efficiency of the proposed algorithm in comparison with the previous researches.

3.2.2. Algorithm Semantic annotator algorithm (Algorithm 1) can be briefly described as an iterative process of information percolation from seed concepts via local association links. The algorithm takes a concept graph G and a weighted set of seed concepts S as input. In the case of our background knowledge, the vertices are concepts, and the edges are associations. Therefore, the aforementioned pseudo-code uses the letters C and A, respectively. An association is represented as a pair (p, q), where p and q are the source and destination concepts. The proposed random walk model assumes the existence of a random surfer that roams this map by stochastically following local association links until reaching to a stationary distribution.

In the first step, the initial value of each concept must be guessed (Haveliwala, 2003). Extracted concepts by the wikifier algorithm (Milne & Witten, 2013) have been leveraged as explicit seed concepts. This weighted list of Wikipedia concepts has been employed as the initial seed distribution of semantic annotator algorithm (S). Each iteration is started by updating the result activation vector, $R^{(t+1)}$, and accumulating $\lambda / |R^{(t+1)}|$ in each entry. This is the probability of landing at any activated concept because of a random jump. It makes the random surfer model an ergodic Markov chain, which guarantees that the iteration calculation of the algorithm will converge (Haveliwala, 2003). Because the random surfer has to be biased towards walking around the initial distribution, seed concepts are promoted at each iteration.

Algorithm 1 Semantic Annotator: Conserved Personalized PageRank

```
procedure CPPR($G, S$)
$(C, A) \leftarrow G$ {Split the conceptual network into "Concepts" and "Associations"}
$R^{(t)} \leftarrow \varnothing$ {The current dynamic activation vector estimate}
$R^{(t+1)} \leftarrow \varnothing$ {The resulting better activation vector estimate}
for all $S_i \in S$ do
    $R_i^{(t+1)} \leftarrow S_i$
end for
repeat
    $R^{(t)} \leftarrow R^{(t+1)}$ {Update the current activation vector with new estimation}
    for all $R_i^{(t+1)} \in R^{(t+1)}$ do
        $R_i^{(t+1)} \leftarrow R_i^{(t+1)} + \lambda / |R^{(t+1)}|$ {Each activated concept has a $\lambda / |R^{(t+1)}|$ chance of random selection}
    end for
    for all $S_i \in S$ do
        $R_i^{(t+1)} \leftarrow R_i^{(t+1)} + S_i \times \lambda / |S|$ {Bias the result to the initial distribution ($S$)}
    end for
    for all $p \in R^{(t)}$ do
        $Q_l \leftarrow$ the set of local concepts such that $(p, q) \in A$ and $q \in C$
        $income \leftarrow (1 - \lambda) R_p^{(t)} / |Q_l|$
        for all $q \in Q_l$ do
            $R_q^{(t+1)} \leftarrow R_q^{(t+1)} + income$
        end for
        $R_p^{(t+1)} \leftarrow R_p^{(t+1)} - income \times |Q_l|$
    end for
until $R$ has not converged
```

The next step is computing the probability of landing on a neighbour concept. This probability is computed by iterating over each concept in the current activation vector, $R^{(t)}$, and retrieving the estimated probability of having reached that concept, $R_p^{(t)}$. From that concept, the random surfer has a λ chance of jumping randomly to an activated concept, or $1 - \lambda$ to spread its weight to a neighbour. There are $|Q_l|$ associations to choose from, so the probability of jumping to a concept $q \in Q_l$ is $income = (1 - \lambda) R_p^{(t+1)} / |Q_l|$, which is accumulated to $R_q^{(t+1)}$. It defines the probability of finding the random surfer at concept q as the sum of all ways; it could be reached from any other concept in the previous pulse.

As described in Section 3.2.1, personalized PageRank (Haveliwala, 2003) converges to general concepts regardless of selected initial seed concepts. Green measure (Ollivier & Senellart, 2007) overcomes this problem by penalizing each concept according to its centrality. Based on the original Green algorithm, PageRank value of each node has been used to determine global centrality of the nodes in the network. Therefore, the main problem in employing the Green algorithm will be its high complexity due to determining PageRank values that require exponential time complexity. This is a prohibitive feature in real-world information-retrieval applications. Based on the proposed algorithm, 'degree centrality' (Steyvers & Tenenbaum, 2005) has been used as a simpler measure for global centrality instead of using PageRank values in the original Green algorithm. The previous researches (Steyvers & Tenenbaum, 2005) have shown that degree centrality measure is a good estimation of global centrality of the nodes in the network in addition to its simplicity. Therefore, in order to limit the contribution of general concepts in the interpretation mechanism, each node is penalized after propagation according to the number of its neighbours (line 21 in Algorithm 1). Noteworthy, it is related to the penalize vector (v in equation (2)) described in Green measure (Ollivier & Senellart, 2007).

The final weight of the concept, c_i represents the proportion of time the random surfer spends visiting it after a sufficiently long time and corresponds to the structural scheme of the association map. The walk terminates when the proportion of time that the random surfer visits each node converges to a stationary distribution. Experimental results show that $R^{(t)}$ exponentially converges to its unique stationary distribution $R^{(\infty)}$ in a few steps (less than four iterations in average). Two convergence criteria are used in this paper. $\varepsilon-$ *truncated* is defined as the divergence of the generated feature vectors ($|R^{(t+1)}-R^{(t)}|<\varepsilon$). Moreover, in order to maintain the computation time within acceptable limits, the iterations of Algorithm 1 can be truncated after T steps ($T-$ *truncated*). Therefore, the iteration of the algorithm will be terminated when the divergence between the generated feature vectors is less than ε or the maximum number of iterations (T) is reached.

3.2.3. Approximations: $T-$ truncated and $\varepsilon-$ truncated convergence criteria One of the most widely used ways to solve random walk issue as described in Algorithm 1 is the iterative method, until the $|R^{(t+1)}-R^{(t)}|$ norm of successive estimation of $R^{(t)}$ is below our threshold (ε), or a maximum iteration step T is reached. Moreover, the number of iteration of the algorithm in a special execution depends on the initial nodes. For some of the initial nodes, the algorithm rapidly converges, while for some other nodes, it may need more iterations to reach to the convergence point. On the other hand, estimating exact values of parameters T and ε is not necessary for each initial node because these parameters only control the estimation error. Empirical experiments show that changes of estimation error are very negligible after some iterations of the algorithm. Therefore, determining upper/lower bound of these parameters will cause acceptable accuracy in addition to reduction of calculation time. In this section, average upper/lower bound of these two parameters is estimated independent of initial seed concepts (S).

Estimation of the average upper/lower bound of the parameters T and ε is independent of the initial nodes (S) and of course is a function of the graph structure. On the other hand, it is not possible to apply random walks for all nodes of the graph in terms of calculation limits; therefore, they must be sampled. To achieve this aim, a set S of 1000 nodes is randomly chosen out of approximately 1 million nodes in the knowledge base, and feature vector for each $s_i \in S$ is generated employing Algorithm 1. Given that S is a large random sample, the evolution of ε with T is considered as a representative of the evolution of an average feature vector. The average upper bound of parameters T and ε is determined using this sample size.

Figure 3 monitors average estimation error ($|R^{(t+1)}-R^{(t)}|$) in every iteration of the algorithm for different values of λ. Each point in Figure 3 presents average estimation error for all samples in a specified iteration of the algorithm (horizontal axis) and a specified value of λ, that is, for random set of $S=\{s_1,s_2,\ldots,s_m\}$ including $m=1000$ of random nodes s_i, ε_i^t is the estimation error of s_i in t^{th} iteration of the algorithm and is defined as follows:

$$\varepsilon_i^t = \left| R_{(s_i)}^{(t+1)} - R_{(s_i)}^{(t)} \right| \tag{3}$$

Finally, the average estimation error in iteration t is defined as follows, which indicates point (t, ε^t) in Figure 3:

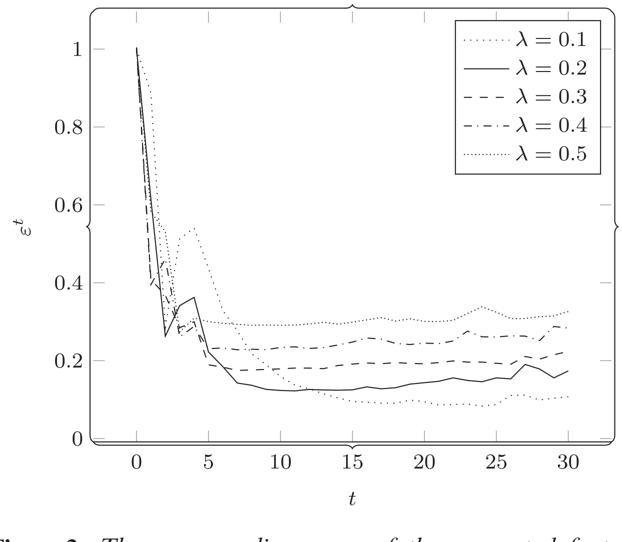


Figure 3: *The average divergence of the generated feature vectors* (ε^t) depending on the number of iterations (t) over the knowledge base, for λ varying from 0.1 to 0.5. The shape of the curves indicates the value of T leading to an acceptable approximation: $\lambda=0.2$ and $T=10$ were chosen in the subsequent experiments.

$$\varepsilon^t = \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i^t \qquad (4)$$

When the number of iterations (t) is increased to infinity, the divergence $\left(\varepsilon_i^t\right)$ is decreased, while the computation time is increased linearly with t. Consequently, $T = 10$ and $\varepsilon = 0.04$ for $\lambda = 0.2$ are chosen as an equilibrium between computation time and accuracy. These parameters have been used in the following experiments of document clustering and classification.

To show tolerance of changes in the estimated error of the selective samples in each point of Figure 3, estimation error distribution of all samples in every iteration of the algorithm for $\lambda = 0.2$ is illustrated in Figure 4. As it is evident in the diagram, sample error distribution is unstable in the initial iterations of the algorithm ($t < 7$), which indicates non-convergence of the algorithm in the initial iterations. As it is clear, after 10 first iterations, the full convergence with specified and partial dispersion can be seen in the sample estimation direction. On the other hand, this experiment shows that estimation error $\left(\varepsilon_i^t\right)$ converges in t^{th} iteration ($t > 10$) of the algorithm for each selective node (s_i) and the divergence will be almost a fixed value.

4. Using semantic annotator for text clustering and classification

Feature generation for a piece of text is a 'one-off' task, which can be applied to the indexing phase. The generated features are leveraged in different applications of NLP. The benefits of using conceptual features are evaluated in two specific applications: document clustering and classification. The main goal of the experiments of this section is presenting the benefits priority of using the Wikipedia concepts proposed by our method rather than words appearing in text (Manning *et al.*, 2008; Szyma & Ski, 2014) or latent features (Deerwester *et al.*, 1990; Landauer & Dumais, 1997). Therefore, experiments have been performed in the field of clustering and classification using standard corpora and algorithms (Szyma & Ski, 2014).

The previous research (Moschitti & Basili, 2004) has shown that efficiency of a text classifier has direct relation to the selective features. Because the main goal of this paper is presenting a method to generate feature, explicit experimentation on feature selection techniques has been avoided as is outside the focus of the paper.

The knowledge of the proposed system comes entirely from the associative network (Section 3.1) and the semantic annotator algorithm (Section 3.2.2). A central claim of this paper is that conceptual features generated by the semantic annotator are not only more informative than words but also wikifier's concepts. Hence, the BOW and wikifier models are used as a baseline in the following experiments.

The generated features for a piece of text are a bag of Wikipedia concepts that are limited to the domain-specific nouns. Furthermore, previous researches (Hu *et al.*, 2008; Gabrilovich & Markovitch, 2009; Hu *et al.*, 2009c) emphasized on the importance of word features in the task of text classification. Therefore, word features are augmented using new conceptual-generated features, the so-called 'enriched BOW' in the following experiments. The result is an enriched feature vector that contains both words and concepts.

4.1. Test collection

There are different corpora for evaluation of the text classification or clustering methods, and each one of them has its own characteristics. In this section, it has been tried to select a set of the documents that have enough diversity in terms of length of document, type of document and

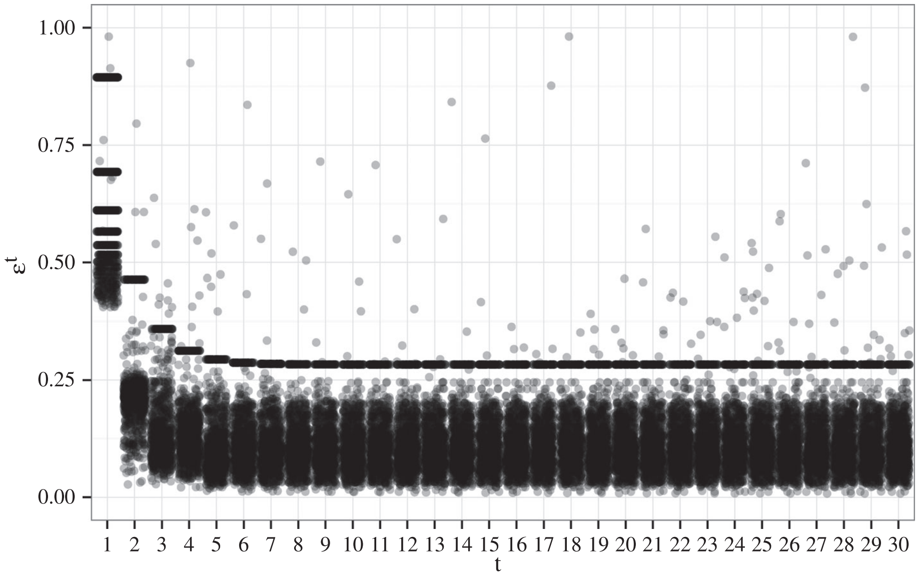


Figure 4: *Estimation error distribution of the selective samples in every iteration of Algorithm 1 ($\lambda = 0.2$).*

number of classes. On this basis, experiments are conducted on a subset of OHSUMED (including specialized documents with short length), 20 NewsGroups (including many training data with variable length) and Reuters (including general documents with variable length) collections (Gabrilovich & Markovitch, 2009; Huang *et al.*, 2012). Each corpus has different properties, topic domains and difficulty levels. The following datasets allow us to comprehensively evaluate the performance of the proposed approach. A brief description about these datasets is presented in the following:

1. Reuters-21578: It is the most often used general purpose dataset in the task of text classification. The original collection containing 11367 manually labelled documents that are classified into 82 categories, with 9494 documents uniquely labelled. Two datasets are created from this corpus:

   (a) Reuters-30: For consistency with previous studies on this collection following previous researches (Hu *et al.*, 2008; Huang *et al.*, 2012), categories with less than 15 documents or more than 200 documents are removed, leaving 30 categories comprising 1658 documents.
   (b) Reuters-90: Following another common practice (Gabrilovich & Markovitch, 2009; Wang *et al.*, 2009), we used the ModApte split (9603 training and 3299 testing documents), and 90 categories with at least one training example and one testing example.
2. Med100: The original OHSUMED collection contains 348566 medical documents that each document is labelled with an average of 13 mesh categories (out of total 14000). Furthermore, each document contains a title, and about two-thirds of them also contain an abstract.

   Following previous researches (Hu *et al.*, 2008; Huang *et al.*, 2012), we choose a subset of 18302 single-label *abstracts* that are classified into 23 categories where with each category contains from 56 to 2876 abstracts.
3. 20NG: This is a well-balanced collection that contains 19997 postings to 20 news groups, comprising 20 categories with about 1000 documents per category. Following previous research (Huang *et al.*, 2012), in order to understand whether topic separation impacts the performance of the proposed method, two datasets are created from this corpus (Huang *et al.*, 2012):
   (a) NewsSim3: This dataset contains closely related topics from 20NG collection ('comp.windows.x', 'comp.graphics' and 'comp.os.ms-windows.misc' categories). It is composed of three closely related categories and 1000 documents per category.
   (b) NewsDiff3: On the opposite of NewsSim3, NewsDiff3 contains three clearly separated categories ('sci.space', 'alt.atheism' and 'rec.sport. baseball') with similar statistics. Considering the subtle distinction between the topics that NewsSim3 contains, it must be more difficult than NewsDiff3 to be clustered or classified.

4.2. Experimental setting

As baseline, the words appearing in the text have been used as feature in text classification and clustering applications. All terms in each document have been first extracted. Then, stop words and all terms that have appeared in the corpus for less than five times have been removed. Finally, the remaining terms have been returned to their main stems with Porter algorithm (Manning *et al.*, 2008). In the following, importance of the terms in each document has been determined using Term Frequency–Inverse Document Frequency (TFIDF) weighing algorithm (Manning *et al.*, 2008). Ultimately, term-document matrix (D) including documents and all terms appearing in them has been formed, and values of this matrix have been normalized based on length of document (Manning *et al.*, 2008). This baseline execution is specified as 'BOW' in our experiments.

To compare the proposed method with features generated using LSA (Deerwester *et al.*, 1990; Landauer & Dumais, 1997), the produced term-document matrix (D) has been transformed to a smaller matrix including k eigenvalue for each corpus with singular-value decomposition. k parameter controls combination of different terms for making latent features. For optimal estimation of k value in each dataset, 95% of the total singular values have been used as the number of latent features (default settings in WEKA (Bouckaert *et al.*, 2010)).[4] This execution is specified as 'LSA' in the following experiments.

It should be noted that reduction of dimensions of term-document matrix has acceptable results only in dealing with many documents. On this basis, this model is almost made based on an external corpus like Wikipedia, and then the generated transformation matrix is used for production of latent features from different documents (Stefanescu *et al.*, 2014). In addition, more efficient latent features can be produced by training model on a specialized corpus relating to the test dataset (such as 'OHSUMED' corpus for 'Med100' dataset). Because a wide range of corpora was employed, it was not possible to train each dataset according to its original corpus. It was preferred to leverage the entire test dataset (for example, all 18302 documents in Med100 dataset) to build latent features that are known as a standard baseline of LSA in previous researches (Deerwester *et al.*, 1990; Landauer & Dumais, 1997). As a result, the obtained accuracy in some datasets would be lower than the expected limit. For example, whereas

[4] This configuration leads to 790, 1394, 1555, 882 and 999 latent features for Reuters-30, Reuters-90, Med100, NewsSim3 and NewsDiff3, respectively.

'Med100' dataset includes small set of short and specialized documents, accuracy of latent features in this dataset is lower than the other datasets including a large set of long and general documents.

At the beginning, conceptual features are generated for each document of the whole dataset. A converter has been developed to import these knowledge-based features into WEKA (Bouckaert *et al.*, 2010), a freely available machine-learning software. To focus our investigation on generating the features that represent the thematic content of a text rather than the clustering or classification method, two standard algorithms with default parameter setting were employed. Furthermore, we used $\lambda = 0.2$, $\varepsilon = 0.04$ and $T = 10$, whose estimation was discussed that estimated in Section 3.2.3.

Text classification has been performed using support vector machine (SVM) classifier (Sebastiani, 2002) that has implemented in LibSVM (Chang & Lin, 2011) package. SVM has been introduced in different papers as pioneering classifier for text classification problem (Sebastiani, 2002; Huang *et al.*, 2012; Szyma & Ski, 2014). The effectiveness of SVM depends on the selection of kernel, the kernel's parameters and soft margin parameter C (Sebastiani, 2002). A common choice in text classification is a simple linear kernel and a constant initial value (default = 1) for the cost parameter C. LibSVM (Chang & Lin, 2011) leverages cross-validation to find the best value of the parameters and uses these values to train the whole training set. Accordingly, 10-fold cross-validation was performed, and paired *t*-test was used to assess the significance. The effectiveness of the proposed representation model is measured in terms of F_1 measure that is composed of the classic notions of precision (π) and recall (ρ) (Sebastiani, 2002; Huang *et al.*, 2012).

Clustering is performed using hierarchical agglomerative clustering with group-average-link (Manning *et al.*, 2008) method over the entire datasets without using any part of them as training set. The effectiveness of a clustering algorithm is measured in terms of internal or external metrics (Aliguliyev, 2009). Internal metrics assess the clusters against their own structural properties and used when there is no knowledge of the real clustering. On the other hand, external metrics refer to comparing a clustering solution with a true clustering and are used when a gold standard dataset is available. Among various external metrics of cluster validation, recent papers (Hu *et al.*, 2009c; Huang *et al.*, 2012) leveraged normalized mutual information. It is a comprehensive measure that reflects the true effectiveness of a clustering algorithm (Manning *et al.*, 2008; Hu *et al.*, 2009c; Huang *et al.*, 2012). So the effectiveness of the proposed representation model is measured in terms of goodness of fit with the existing categories in the dataset using the normalized mutual information measure (Hu *et al.*, 2009c; Huang *et al.*, 2012). For each dataset, the number of clusters is considered equal to the number of categories. Each cluster is labelled with the category that is the most frequent one in that cluster. Therefore, a text is correctly clustered if the cluster it is assigned to is labelled with the category it belongs to. Let Φ and Ω denote the set of clusters and categories, respectively. The normalized mutual information measure is defined as follows:

$$NMI(\Phi, \Omega) = \frac{I(\Phi; \Omega)}{[H(\Phi) + \mathrm{H}(\Omega)]/2} \tag{5}$$

where I is the mutual information between the set of clusters and the set of categories and H is the entropy (Huang *et al.*, 2012).

4.3. Experimental results

The effectiveness of the proposed approach in comparison with the baseline methods (BOW and wikifier) and previous statistical (Landauer & Dumais, 1997) and knowledge-based (Huang *et al.*, 2012) approaches is illustrated in Table 1. The different rows of the table correspond to various datasets, as defined in Section 4.1. To our knowledge, the best result for these datasets has been reported by Huang *et al.* (2012). They proposed a supervised document similarity measure that assesses similarity at both the lexical and semantic levels and learns from human judgments how to combine them by using machine-learning techniques.

As earlier studies found, most BOW features are indeed useful for SVM text classification (Gabrilovich & Markovitch, 2009). On the other hand, enriching the BOW with new generated conceptual features achieves significant improvement on most corpora, especially Med100 dataset.

Table 1: *The effect of feature generation in the task of text classification in comparison with previous knowledge-based (Huang et al., 2012) and statistical (Landauer & Dumais, 1997) methods*

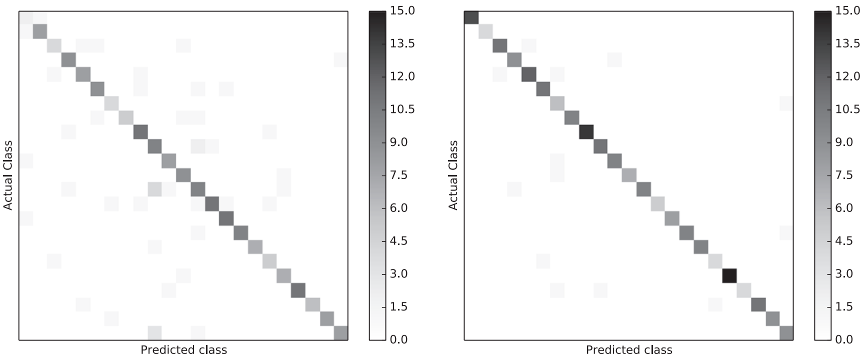
	Baseline		*Previous works*		*Proposed method*
	BOW	*Wikifier*	*(Huang et al., 2012)*	*LSI (Landauer & Dumais, 1997)*	*Enriched BOW*
Reuters-30	0.902	0.830	0.924	0.309	0.926
Reuters-90	0.719	0.603	–	0.421	0.723
Med100	0.744	0.601	0.591	0.766	0.956
NewsSim3	0.960	0.907	0.833	0.955	0.966
NewsDiff3	0.999	0.972	0.976	0.998	1.000

BOW, Bag of Words; LSI, Latent Semantic Indexing.

As described in Section 4.1, Med100 contains domain-specific documents that are perfectly covered by Wikipedia's concepts. In order to show efficiency of the proposed approach, the empirical results obtained on Med100 dataset are presented in Figure 5 as a confusion matrix over individual classes. The diagonal of the confusion matrix represents the absolute accuracy of the corresponding class.

As mentioned earlier, NewsSim3 and NewsDiff3 contain a lot of training examples. Therefore, the effectiveness of BOW is satisfactory. To investigate the impact of the conceptual features on these datasets, the proportion of training documents is varied from 5% to 95% in increments of 5%, and the remaining examples are used for testing. The results are illustrated in Figure 6. The important observation is that when the training set is small, the enriched feature space significantly outperforms the baseline BOW. It must be noted that whereas obtaining labelled training data is often expensive in different applications, using small training set can have important effect in practice.

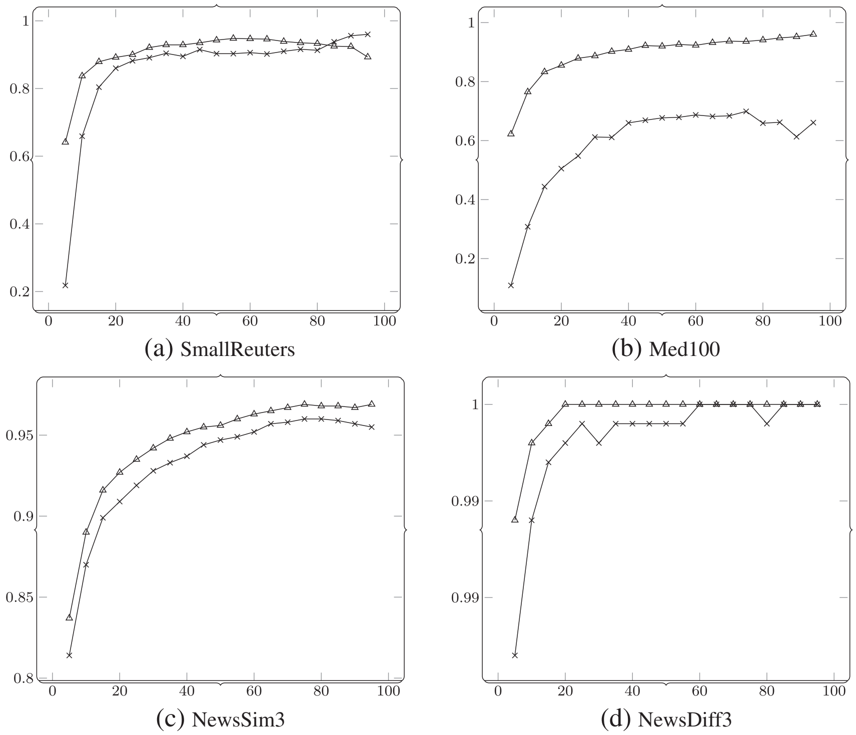
The use of the conceptual features does not lead to considerable improvement in comparison with the BOW model in any of the two datasets Reuters-90 and Reuters-30. On the other hand, the same method on Med100 dataset from the OHSUMED corpus considerably improves quality



(a) Bag of words

(b) Enriched bag of words

Figure 5: *Confusion matrix of support vector machine classifier over Med100 dataset. Horizontal axis represents the instances in a predicted class, while vertical axis represents the instances in an actual class.*



(a) SmallReuters

(b) Med100

(c) NewsSim3

(d) NewsDiff3

Figure 6: *Learning curves for different sizes of the training set over different datasets. The proportion of training examples is varied from 5 to 95 in the horizontal axis, and the weighted F measure is shown in the vertical axis. The remaining examples are used for testing.* —×— *and* —△— *indicate bag of words (BOW) and the enriched BOW, respectively.*

of classification. It is due to the type of the documents in Reuters-21578 corpus from which both datasets have been derived. The documents available in the Reuters-21578 corpus are almost long and general documents. On the contrary, documents of the OHSUMED corpus are short and contain specialized words in the field of medicine. These words almost have different forms, and the set of their synonymous words is also very rich. Under these conditions, the words appearing in each document are not only enough for description of content of that document (due to short length of the documents) but also mismatch problem is inevitable when matching terms of different documents (due to specialized nature of terms and broad set of the synonymous words for each term). In other words, a set of the performed experiments concludes that the approach used in the proposed method (enriching BOW model using concepts) leads to considerable improvement in efficiency of the classifier especially in dealing with short and specialized documents. It is compatible with the results obtained in the recent papers in the field of enriching BOW model (Hu *et al.*, 2009b; Wang *et al.*, 2009; Ferragina, 2011).

The results of the clustering algorithm in comparison with the previous statistical (Landauer & Dumais, 1997) and knowledge-based (Huang *et al.*, 2012) approaches are presented in Table 2. Noteworthy, word features provide useful information in general-purpose datasets ('Reuters-30', 'NewsSim3' and 'NewsDiff3'), while wikifier (Milne & Witten, 2013) is dominant in the domain-specific dataset ('Med100'). Furthermore, knowledge-based features significantly improve the prediction accuracy, especially when evaluating on hard datasets ('Med100' and 'NewsSim3') where clearly separated features are required.

4.4. Discussion

In this section, the applicability of the proposed method is demonstrated in the task of conceptual feature generation using a sample document, and then the advantages and disadvantages of the proposed method are discussed comprehensively. Whereas using words as features (BOW) is a well-known classic representation of text documents, this representation was employed as a baseline of our experiments (the so-called 'BOW' in Tables 1 and 2). For example, consider the following sample document from OHSUMED collection:

> 'Maternal hemodynamics in normal and preeclamptic pregnancies: a longitudinal study Preeclampsia is a disease unique to pregnancy that contributes substantially to maternal and fetal morbidity and mortality. The condition has been thought to be one of hypoperfusion in which increased vascular resistance characterizes the associated hypertension. This study was designed to test an alternative hypothesis, that preeclampsia is characterized by high cardiac output. In a blinded longitudinal study of nulliparas with uncomplicated pregnancies, cardiac output was measured serially by Doppler technique. Cardiac output was elevated throughout pregnancy in patients who became preeclamptic (P = .006). Six weeks post-partum, the hypertension of the preeclamptic subjects had resolved but cardiac output remained elevated (P = .001) and peripheral resistance remained lower than in the normotensive subjects (P = .001). This study demonstrates that preeclampsia is not a disease of systemic hypoperfusion and challenges most current models of the disease based on that assumption.'

As described in Section 4.2, after removing stop words and all words that have appeared in the corpus for less than five times, the remaining words have been returned to their main stems. Table 3 shows the most important features in the corresponding feature vector using BOW representation model. The aforementioned sample document is about the

Table 2: *The effect of feature generation in the task of text clustering in comparison with previous knowledge-based (Huang et al., 2012) and statistical (Landauer & Dumais, 1997) studies using hierarchical agglomerative clustering with group average link*

	Baseline		*Previous works*		*Proposed method*
	BOW	*Wikifier*	*(Huang et al., 2012)*	*LSI (Landauer & Dumais,1997)*	*Enriched BOW*
Reuters-30	0.592	0.532	0.696	0.227	0.620
Med100	0.299	0.317	0.365	0.109	0.406
NewsSim3	0.067	0.041	0.176	0.014	0.312
NewsDiff3	0.509	0.389	0.613	0.026	0.609

BOW, bag of words.

Table 3: *Top word features using bag of words representation model for a sample document of OHSUMED collection*

	Word features									
	Pre-eclampt	*Nullipar*	*Hypoperfus*	*Post-partum*	*Cardiac*	*Pre-eclamps*	*Uncomplic*	*Normotens*	*Longitudin*	*Vascl*
Weight	6.2	5.6	5.6	5.1	4.9	4.6	4.5	4.1	3.7	3.6

impact of cardiac output on pre-eclampsia and belongs to ‘female diseases and pregnancy complications’ class. As it is clear, by using words as features, the predicted class is ‘cardiovascular diseases’. The outcome is predictable because it contains features such as ‘vascl’ and ‘cardiac’, which are very important features in ‘cardiovascular diseases’ class. On the other hand, in the field of conceptual feature generation, new emerging conceptual features that represent the topical semantic of the input document are desirable. In the following, it is demonstrated how the aforementioned sample document is represented using the proposed semantic annotator.

As presented in Figure 1 in Section 3, in order to use the concept network for feature generation, the first issue is mapping each text document to a set of concepts in the network. Wikifier (Milne & Witten, 2013) was employed for mapping text fragments to their corresponding concepts in the network. The most important concepts extracted by the wikifier algorithm (Milne & Witten, 2013) are shown in Table 4 with the corresponding weights. These concepts are called ‘seed concepts’ shown as S in Algorithm 1. Although it is possible to represent each document with seed concepts (‘wikifier’ column in Tables 1 and 2), semantic annotator (Algorithm 1) tried to extend the topical semantics of the initial seed concepts by finding new emergent concepts describing the gist of the document. Algorithm 1 gets the seed concepts (S) and ranks all concepts in the conceptual network (G) according to their relevance to the initial seed concepts.

The generated features using semantic annotator for the aforementioned sample document are presented in Table 4. New emerging features (such as ‘obstetrics’) completely represent the context of the input text and have great impact on learning a classifier. Whereas these features do not exist in the document, wikifier is unable to detect them. On the other hand, semantic annotator (Algorithm 1) gets the initial seed concepts (S) and generates new emerging concepts with the aid of the underlying conceptual network (G).

The nature of the features generated by the proposed algorithm is completely different with word features. Whereas features are Wikipedia concepts, the generated features are highly limited to the domain-specific nouns. On the other hand, word features cover a broad range of nouns, verbs and adjectives. Previous researches (Hu *et al.*, 2008; Gabrilovich & Markovitch, 2009; Hu *et al.*, 2009c) emphasized on the importance of word features in the task of text classification and clustering. Therefore, word features (Table 3) are enriched with new conceptual features (Table 4), the so-called ‘enriched BOW’ in our experiments (Tables 1 and 2). The result is a feature vector containing both word and conceptual features. It should be noted that word features have been normalized before the augmentation of conceptual features.

The features generated by wikifier (Milne & Witten, 2013) or LSA (Landauer & Dumais, 1997) made explicit words in the document. Therefore, they have become redundant when augmenting these features with words explicitly presented in the document (BOW features). Wikifier features presented in Table 4 have a corresponding word in the document. For example, the corresponding word for ‘disease’ feature is ‘disease’ word presented frequently in the input text. It is clear that adding these features is not informative in the representation of the input text.

On the other hand, the proposed conceptual features are Wikipedia concepts, which have a topical similarity with the words presented in the document. For example, consider again the generated features presented in Table 4. As mentioned before, new emerging features (such as ‘obstetrics’) completely represent the context of the input text. This emerging feature does not have a corresponding word in the document. Therefore, combining the words presented in the document with these new emerging conceptual features is a promising approach to better represent the input document. Table 5 shows the result of enriching BOW representation model with wikifier (Milne

Table 5: *Enriching bag of words representation model (‘BOW’) with wikifier (Milne & Witten, 2013) (‘BOW + wikifier’), LSA (Landauer & Dumais, 1997) (‘BOW + LSA’) and the proposed feature generation method (‘enriched BOW’) in the tasks of document classification and clustering using Med100 dataset*

	Classification			*Clustering*
	π	ρ	F_1	*NMI*
BOW	0.866	0.680	0.744	0.299
BOW + wikifier	0.852	0.680	0.738	0.308
BOW + LSA	0.866	0.680	0.744	0.316
Enriched BOW	0.945	0.970	0.956	0.406

LSA, latent semantic analysis; NMI, Normalized Mutual Information.

Table 4: *Top conceptual features generated by the proposed semantic annotator using Algorithm 1*

	Conceptual features									
	Disease	*Circulatory system*	*Obstetrics*	*Pregnancy*	*Hypertension*	*Pre-eclampsia*	*Cardiac output*	*Shock (medical)*	*Death*	*Childbirth*
Weight	1.0	0.63	0.57	0.40	0.37	0.31	0.25	0.21	0.13	0.08
seed concepts	√			√	√	√	√	√	√	

Each feature corresponds to a Wikipedia article. Seed concepts are shown with √ notation.

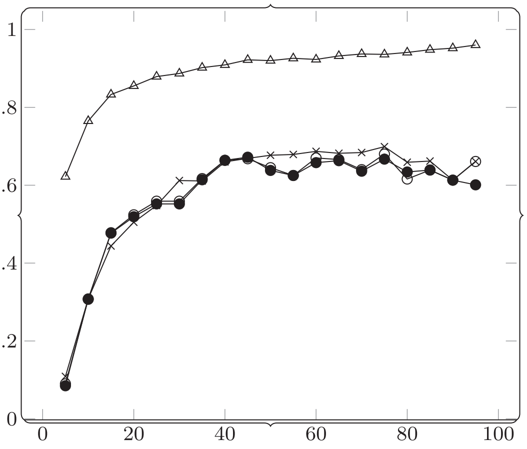


Figure 7: *Learning curve for different sizes of the training set (horizontal axis) over Med100 dataset. The remaining examples are used for testing. —×—, —●—, —○— and —△— indicate bag of words (BOW), BOW + wikifier, BOW + latent semantic analysis and the proposed enriched BOW, respectively.*

& Witten, 2013), LSA (Landauer & Dumais, 1997) and the proposed feature generation method in the tasks of document classification and clustering using Med100 dataset. Other datasets produced nearly the same result.

Learning curve for different sizes of the training set over Med100 dataset has been presented in Figure 7. The proportion of training examples is varied from 5 to 95 in the horizontal axis, and the weighted F measure is shown in the vertical axis. The remaining examples are used for testing. It is clear that combining new generated features of LSA (Landauer & Dumais, 1997) or wikifier (Medelyan *et al.*, 2009) to BOW representation model does not contribute to the classifier accuracy. On the other hand, combining the proposed conceptual features to BOW representation model significantly improves the performance. Other datasets produced nearly the same result.

5. Conclusion and future work

Conceptual feature generation for textual information is a fundamental issue in the field of NLP and concept-based information retrieval. In this paper, a knowledge-based feature generator for textual information is proposed. It leverages a hyperlinked encyclopaedia (such as Wikipedia in our experiments) as a valuable source of common-sense knowledge. In addition, the semantic annotator embedded in the proposed model is responsible for enriching the input text. This model leads to concepts that cannot be deduced from the input text alone. Consequently, the proposed model outperforms the conventional BOW (Jurgens & Stevens, 2010) as well as statistical latent representation models (Landauer & Dumais, 1997). Experimental results confirmed the significance of the presented model in two fundamental tasks of NLP especially in dealing with short and specialized documents.

Over the last decades, the explosion of applications in the field of social media, search engine advertising and instant messaging introduced a new challenging scenario where texts are very short and poorly written (Ferragina, 2011; Rosso *et al.*, 2013). Whereas these texts do not provide enough word co-occurrence or shared context, the performance of models based on word features becomes quite limited in this new scenario. Our experiments revealed that using the proposed–enriched BOW representation model is promising in these new challenging applications.

On the other hand, the proposed approach is applicable to many NLP tasks whose input is textual information and the output is a decision based on its context. Text summarization, information retrieval and query understanding (Hu *et al.*, 2009a) are promising applications. Furthermore, employing Wikipedia as background knowledge is a great potential of the proposed model that facilitates future applications. For example, according to the Wikipedia writing rules (Medelyan *et al.*, 2009), the first sentence of each article concisely describes the content of the corresponding article. It is used as a textual definition that can be leveraged in automatic summarization and glossary building. Furthermore, Wikipedia has been admired for its coverage of name entities and specialized concepts, while it has been criticized for lack of coverage for general terms (Gurevych & Wolf, 2010; Hovy *et al.*, 2013). Leveraging recent hybrid multilingual knowledge sources (Navigli & Ponzetto, 2012; Nastase & Strube, 2013) can be beneficial and improve the overall effectiveness of the proposed system.

Although enriching the BOW with conceptual features is a prominent approach, increasing the dimensionality is inevitable. ‘Curse of dimensionality’ is a core problem in the task of clustering and classification. We believe that the most interesting benefit of the proposed conceptual annotator is the structured knowledge attached to the input text document. By employing the links between the conceptual features (Elci, 2011; Jadidinejad *et al.*, 2015), it is possible to cluster relevant or redundant features and provide more informative feature space.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

Agichtein, E., E. Gabrilovich and H. Zha (2009) The social future of web search: modeling, exploiting, and searching collaboratively generated content, *IEEE Data Engineering Bulletin*, **32**(2), 52–61.

Aliguliyev, R.M. (2009) Performance evaluation of density-based clustering methods, *Information Sciences*, **179**(20), 3583–3602.

Anderka, M. and B. Stein (2009) The ESA retrieval model revisited. In *Proceedings of the 32nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '09, ACM, New York, NY, USA, 670–671.

Armstrong, T.G., A. Moffat, W. Webber and J. Zobel (2009) Has ad hoc retrieval improved since 1994?. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, ACM, New York, NY, USA, 692–693.

Bouckaert, R.R., E. Frank, M.A. Hall, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten (2010) WEKA – experiences with a Java open-source project, *Journal of Machine Learning Research*, **11**, 2533–2541.

Chang, C.C. and C.J. Lin (2011) LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, **2**(27), 1–27 27.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman (1990) Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41**(6), 391–407.

Egozi, O., S. Markovitch and E. Gabrilovich (2011) Concept-based information retrieval using explicit semantic analysis, *ACM Transactions on Information Systems*, **29**(28), 1–8 34.

Elci, A. (2011) Article: text classification by PNN-based term re-weighting, *International Journal of Computer Applications*, **29**(12), 7–13.

Ferragina, P. (2011) Beyond the bag-of-words paradigm to enhance information retrieval applications. In *Proceedings of the Fourth International Conference on Similarity Search and Applications, SISAP '11*, ACM, New York, NY, USA, 3–4.

Fodeh, S., B. Punch and P.N. Tan (2011) On ontology-driven document clustering using core semantic features, *Knowledge and Information Systems*, **28**(2), 395–421.

Gabrilovich, E. and S. Markovitch (2007) Harnessing the expertise of 70,000 human editors: knowledge-based feature generation for text categorization, *Journal of Machine Learning Research*, **8**, 2297–2345.

Gabrilovich, E. and S. Markovitch (2009) Wikipedia-based semantic interpretation for natural language processing, *Journal of Artificial Intelligence Research*, **34**, 443–498.

Gouws, S., G.J. van Rooyen and H.A. Engelbrecht (2010) Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. In *Proceedings of the 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, Coling 2010 Organizing Committee, Beijing, China, 46–54.

Gurevych, I. and E. Wolf (2010) Expert-built and collaboratively constructed lexical semantic resources, *Language and Linguistics Compass*, **4**(11), 1074–1090.

Gurevych, I. and T. Zesch (2013) Collective intelligence and language resources: introduction to the special issue on collaboratively constructed language resources, *Language Resources and Evaluation*, **47**(1), 1–7.

Haveliwala, T.H. (2003) Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search, *IEEE Transactions on Knowledge and Data Engineering*, **15**, 784–796.

Hovy, E., R. Navigli and S.P. Ponzetto (2013) Collaboratively built semi-structured content and artificial intelligence: the story so far, *Artificial Intelligence*, **194**, 2–27.

Hu, J., L. Fang, Y. Cao, H.J. Zeng, H. Li, Q. Yang and Z. Chen (2008) Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, ACM, New York, NY, USA, 179–186.

Hu, J., G. Wang, F. Lochovsky, J.T. Sun and Z. Chen (2009a) Understanding user's query intent with Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, ACM, New York, NY, USA, 471–480.

Hu, X., N. Sun, C. Zhang and T.S. Chua (2009b) Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, ACM, New York, NY, USA, 919–928.

Hu, X., X. Zhang, C. Lu, E.K. Park and X. Zhou (2009c) Exploiting Wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, ACM, New York, NY, USA, 389–396.

Huang, L., D. Milne, E. Frank and I.H. Witten (2012) Learning a concept-based document similarity measure, *Journal of the American Society for Information Science and Technology*, **63**(8), 1593–1608.

Hughes, T. and D. Ramage (2007) Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Association for Computational Linguistics, Prague, Czech Republic, 581–589.

Jadidinejad, A. and F. Mahmoudi (2014) Unsupervised short answer grading using spreading activation over an associative network of concepts/la notation sans surveillance des rÃ©ponses courtes en utilisant la diffusion dâ€™activation dans un rÃ©seau associatif de concepts, *Canadian Journal of Information and Library Science*, **38**(4), 287–303.

Jadidinejad, A.H., F. Mahmoudi and M.R. Meybodi (2015) Clique-based semantic kernel with application to semantic relatedness, *Natural Language Engineering FirstView*, 1–18. http://dx.doi.org/10.1017/S135132491500008X

Jurgens, D. and K. Stevens (2010) The S-space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations, ACLDemos '10*, Association for Computational Linguistics, Stroudsburg, PA, USA, 30–35.

Landauer, T.K. and S.T. Dumais (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review*, **104**(2), 211–240.

Manning, C.D., P. Raghavan and H. Schütze (2008) *Introduction to Information Retrieval*, New York, NY, USA: Cambridge University Press.

Medelyan, O., D. Milne, C. Legg and I.H. Witten (2009) Mining meaning from Wikipedia, *Internationa Journal of Human-Computer Studies*, **67**(9), 716–754.

Milne, D. and I.H. Witten (2013) An open-source toolkit for mining Wikipedia, *Artificial Intelligence*, **194**(0), 222–239.

Moschitti, A. and R. Basili (2004) Complex linguistic features for text classification: a comprehensive study. In McDonald, S. and J. Tait (editors), *Advances in Information Retrieval, Lecture Notes in Computer Science* **2997**, Springer, Berlin Heidelberg, 181–196.

Nasir, J.A., I. Varlamis, A. Karim and G. Tsatsaronis (2013) Semantic smoothing for text clustering, *Knowledge-Based Systems*, **54**(0), 216–229.

Nastase, V. and M. Strube (2013) Transforming Wikipedia into a large scale multilingual concept network, *Artificial Intelligence*, **194**(0), 62–85.

Navigli, R. and S.P. Ponzetto (2012) Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence*, **193**(0), 217–250.

Ollivier, Y. and P. Senellart (2007) Finding related pages using Green measures: an illustration with Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence* **2** AAAI'07, Vancouver, Canada: AAAI Press, 1427–1433.

Ramage, D., A.N. Rafferty and C.D. Manning (2009) Random walks for text semantic similarity. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language*

Processing, TextGraphs-4, Association for Computational Linguistics, Stroudsburg, PA, USA, 23–31.

Rosso, P., M. Errecalde and D. Pinto (2013) Analysis of short texts on the web: introduction to special issue, *Language Resources and Evaluation*, **47**(1), 123–126.

Sahlgren, M. (2006) The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis, Stockholm: Institutionen fÃ¶r lingvistik.

Sebastiani, F. (2002) Machine learning in automated text categorization, *ACM Computing Surveys*, **34**(1), 1–47.

Stefanescu, D., R. Banjade and V. Rus (2014) Latent semantic analysis models on Wikipedia and TASA. In Chair, N.C.C., K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (editors), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14')*, European Language Resources Association (ELRA), Reykjavik, Iceland.

Steyvers, M. and J.B. Tenenbaum (2005) The large-scale structure of semantic networks: statistical analyses and a model of semantic growth, *Cognitive Science*, **29**(1), 41–78.

Syed, Z., T. Finin and A. Joshi (2008) Wikipedia as an ontology for describing documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*, Seattle, Washington: AAAI Press.

Szyma, Å. and J. Ski (2014) Comparative analysis of text representation methods using classification, *Cybernetics and Systems*, **45**(2), 180–199.

Turney, P.D. and P. Pantel (2010) From frequency to meaning: vector space models of semantics, *Journal of Artificial Intelligence Research*, **37**(1), 141–188.

Wang, J. and J. Han (2014) Entity linking with a knowledge base: issues, techniques, and solutions, *IEEE Transactions on Knowledge and Data Engineering*, **99**(PrePrints), 1.

Wang, P. and C. Domeniconi (2008) Building semantic kernels for text classification using Wikipedia. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, ACM, New York, NY, USA, 713–721.

Wang, P., J. Hu, H.J. Zeng and Z. Chen (2009) Using Wikipedia knowledge to improve text classification, *Knowledge and Information Systems*, **19**(3), 265–281.

Yazdani, M. and A. Popescu-Belis (2013) Computing text semantic relatedness using the contents and links of a hypertext encyclopedia, *Artificial Intelligence*, **194**(0), 176–202.

Yeh, E., D. Ramage, C.D. Manning, E. Agirre and A. Soroa (2009) *Wikiwalk: Random Walks on Wikipedia for Semantic Relatedness. In: Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs-4*, Association for Computational Linguistics, Stroudsburg, PA, USA, 41–49.

Zesch, T. and I. Gurevych (2010) Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words, *Natural Language Engineering*, **16**(01), 25–59.

Zhang, Z., A.L. Gentile, L. Xia, J. Iria and S. Chapman (2010) A random graph walk based approach to computing semantic relatedness using knowledge from Wikipedia. In Chair, N.C.C., K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (editors), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, European Language Resources Association (ELRA), Valletta, Malta.

The authors

Amir H. Jadidinejad

Amir H. Jadidinejad is a faculty member of Islamic Azad University of Qazvin (QIAU). His research interests include information retrieval, machine learning and statistical data analysis. He is also interested in applying state-of-the-art models of information retrieval and machine learning to very large collections, such as WWW.

Fariborz Mahmoudi

Fariborz Mahmoudi received his BS, MS and PhD degrees in Computer Engineering. He has been an assistant professor in Computer and IT Engineering faculty at the Qazvin Azad University, Iran. As a researcher too, he was in the Information and Communication Research Center of Iran (Tehran, Iran) between 2002 and 2006, and as a senior researcher in the Mechatronics Research Lab of Qazvin Azad University from 2006 to 2013. Since 2013, he worked in the image analysis lab of Henry Ford Health System (Detroit, MI, USA). His research interests include machine learning, machine vision and information (image and text) retrieval. Dr Mahmoudi has published more than 100 papers in the international scientific journals and conferences proceedings.

M. R. Meybodi

Mohammad Reza Meybodi received the BS and MS degrees in Economics from Shahid Beheshti University in Iran, in 1973 and 1977, respectively. He also received the MS and PhD degrees from Oklahoma University, USA, in 1980 and 1983, respectively, in Computer Science. Currently, he is a full professor in Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran. His research interests include learning systems, parallel algorithms and soft computing.