Autonomous and Adaptive Identification of Topics in Unstructured Text

Louis Massey

Department of Mathematics and Computer Science,
Royal Military College, Kingston, Canada, K7K 7B4
massey@rmc.ca

Abstract. Existing topic identification techniques must tackle an important problem: they depend on human intervention, thus incurring major preparation costs and lacking operational flexibility when facing novelty. To resolve this issue, we propose an adaptable and autonomous algorithm that discovers topics in unstructured text documents. The algorithm is based on principles that differ from existing natural language processing and artificial intelligence techniques. These principles involve the retrieval, activation and decay of general-purpose lexical knowledge, inspired by how the brain may process information when someone reads. The algorithm handles words sequentially in a single document, contrary to the usual corpus-based bag-of-words approach. Empirical results demonstrate the potential of the new algorithm.

Keywords: Text Analysis, Information Retrieval, Knowledge Management, Topics Identification, Text Mining, Text Clustering, Document Classification, Topics Modeling.

1 Introduction

With the continually increasing amount of human knowledge available as electronic text, the design of algorithms capable of determining what a text document is about has been an important research area for years. Despite important progress, computers still have major problems making sense of text. On one hand, the problem originates from the ambiguous nature of text and computers inability to deal with this ambiguity. On the other hand, knowing what a text document is about requires large amounts of difficult and costly to assemble knowledge. The efforts currently deployed to build the Semantic Web illustrate the scale of human intervention required to acquire the knowledge enabling intelligent search and access to text information on the Internet.

We investigate the task of finding the main topics present in a document. The word : topics is taken in its intuitive sense of what a text is about, as identified by a few summarizing and evocative keywords. These keywords are not necessarily present in the text under analysis. We particularly focus on making the topic identification task deal with text ambiguity in an autonomous and adaptive manner. We contribute new fundamental principles that achieve this aim. The principles differ from existing techniques and are inspired by the way humans may be processing information when

they read. We named these principles ReAD, which stands for REtrieval, Activation and Decay. The main idea is that as words are read sequentially, they activate regions of the brain that contain information related to each word. This activation augments as more words accessing the same regions are read, but also decays when regions are rarely accessed. At the end of the document, the strongest activated regions indicate what the text was about.

The innovation stems from the ReAD principles that allow for the handling of text ambiguity to determine topics in an adaptive and autonomous manner. Indeed, the approach we propose avoids both the issues of human intervention and of dependence on inflexible corpus-based training that exist with current techniques. We demonstrate an algorithmic implementation based on the ReAD principles and show experimentally that the algorithm can eliminate the problem of ambiguity and capture the general meaning – the topics - of a text document. The algorithm emulates the ReAD cognitive process using only lexical information stored in a dictionary.

The paper is organized as follows: section 2 contains an overview of related work and identifies the problems with existing topics identification methods. In section 3, we introduce the ReAD principles for topics identification, while in section 4 an algorithmic implementation of these principles is described. In section 5, we report on empirical evaluations of the algorithm and discuss the results obtained.

2 Related Work

The topics of documents are often sought in co-occurrence information found in text corpora. Applications of this idea can be found in methods like Latent Dirichlet Allocation (LDA) [1] and Latent Semantic Analysis (LSA) [2, 3]. The problems with these methods include a dependence on the availability of large domain corpora usually assembled at high cost by humans and a lack of adaptability since topics are determined based on a static snapshot of the data present in the corpus during training.

Other techniques to identify the topics of documents based on a large set of documents are document classification [4, 5] and text clustering [6, 7]. Document classification involves the application of supervised machine learning techniques to settle on a set of rules acquired from the recognition and generalization of patterns in training samples. There are two main problems with classification. First, human experts must initially expand much effort to acquire and prepare a set of training documents labeled with the correct topics. Second, once the classifier has been learned, they often do not fare well in the face of novelty (for example, given a new topic) and must be retrained. Contrary to document classification, text clustering operates without training: it aims at grouping documents based on the similarity of their content. Clustering is hence more autonomous and adaptable than classification, but is generally less reliable [8].

An additional problem with existing topic identification techniques is that documents are usually represented in computer memory as high dimensional vectors using the bag-of-words model of text [9]. The value of each numerical component is based on statistical information obtained from the overall vocabulary from a training text collection, using weighting schemes such as TFIDF [10]. Under this

representation, the compositional construction of meaning from the order of words within a particular document is ignored and terms considered unrelated. However, sentences use word order and related terms to unify text and provide information to the reader about the main topics of the document [11]. Some researchers have investigated the use of knowledge to enrich document vectors semantically [12,13], partly to address this problem. This solution however does not address the critical loss of document word order and furthermore does nothing to resolve issues of low autonomy and adaptability arising from the dependence on a training set to build the bag-of-words representation.

Another approach is to harness natural language processing (NLP) techniques to adequately determine the underlying meaning of text. NLP can also possibly be exploited to improve the effectiveness of classification and clustering [14]. NLP techniques aim at deriving meaning of a particular document by performing, among other steps, syntactic and semantic analysis. The former requires that the proper set of grammatical rules be coded to isolate the syntactic structure of the document sentences. The latter is based on large quantities of encyclopedic knowledge [15] to infer the meaning of text. The preparation of syntactic rules and of encyclopedic knowledge is a costly and lengthy endeavor. For this reason, traditional NLP tends to work only in restricted domains and does not scale well to real-life applications.

Also related to our work is automated semantic tagging and cross-referencing [16, 17]. Again, there is a strong dependence on human intervention and knowledge acquisition. Keyword extraction (also known as term recognition or automatic indexing) [18-20] is another research area that aims at finding topics. In this case, one aims at extracting keywords from documents usually with some form of frequency-based measure of word importance within a document or across a corpus [10, 21-22]. One thus obtains the final product (keywords) for a search or text mining task, or alternatively the keywords extracted can be used as features for further processing such as supervised learning or clustering. Either way, frequency-based information has its limits in its ability to identify the most discriminant features for learning or the most evocative keywords for direct consumption. For this reason, additional sources of evidences are investigated by many researchers (for e.g., [12-13], [23-24]). These include ontologies or other knowledge bases that provide background knowledge to reason about or at least establish basic relationships between words and the concepts they represent; domain corpora for extracting statistical information deemed to carry information on conceptual relationships between words; and, NLP to establish the syntactic structure of the text and determine its semantic nature. Clearly, the same problems of costly human intervention and lack of adaptability arise again.

The specific problem we aim at solving here can be summarized as follows: existing techniques to identify topics depend on human interventions to acquire knowledge and to handcraft training sets. The training set is used to build a bag-of-word representation of documents, based on the assumption that the documents in it are representative of future circumstances. The need to handcraft knowledge makes existing topics identification systems costly to develop, while the dependence on training sets makes them little adaptive to new situations. Our objective is to solve these problems with an approach based on fundamentally different principles than those used by existing methods.

3 The ReAD Principles

To address the issues just described, we propose an approach that differs from traditional keyword extraction, corpus-based co-occurrence analysis, classification, clustering and NLP. The approach is intuitively inspired by how humans may determine the topics of a text document when they read [25, 26]. The general idea is as follows: When one reads a document, it can be expected that each word read sequentially will trigger multiple neural activations corresponding to memory areas of the brain associated with the word in question. Over time, neural activation decays, unless subsequent words incrementally activate the same regions. At the end of the document, the concepts associated with the most activated regions in the reader's brain constitute the topics of the document.

The fundamentally important principles at play are first that words read sequentially cause the retrieval and activation of knowledge items associated with each word. Second, that accrued activation of overlapping memory areas where word-related knowledge is stored accumulates over time, while memory areas infrequently accessed decay. Third, that a convergence onto particular areas will take place as words after words in the text focus onto a few common knowledge items. The hypothesis is that in the end, the interplay of activation and decay will cause just a few areas to be discriminately more activated than others, thus identifying the predominant knowledge items, which can then be interpreted as the main concepts – the topics – present in the text. We call these fundamental principles ReAD for Retreival (of stored knowledge), Activation and Decay.

Assuming usage of existing general-purpose knowledge, an algorithm based on the ReAD principles has the advantage of eliminating many problems common with existing approaches. First, it does not depend on costly and laborious knowledge acquisition efforts. Second, it does not rely on training documents statistics and is able to determine a document's topics in isolation from other documents. Third, the algorithm works in a single pass over the text data, which could benefit real-time applications such as social networks mining and streaming newsfeed analysis (for instance in the context of business intelligence). And lastly, the algorithm abandons the vector representation of documents to take into account word order and the compositional effect of words towards meaning.

4 Algorithmic Implementation

An algorithm based on the ReAD principles is shown in Fig 1. The main steps corresponding to the ReAD principles are as follows: step 3.2.2 is where knowledge related to the word currently under consideration (w_j) is retrieved; steps 3.3.1.1 and 3.3.3.2 are respectively where the activation is incremented and decayed. The knowledge source emulating the brain for retrieval can be any existing source of lexical information (e.g., a dictionary) and doesn't depend on particular representation formalism. The only requirement is that it be able to return a list of words describing or defining the word w_j being queried. Here, as a surrogate for knowledge stored in the human brain, we access WordNet [27], a database that, among other features,

describes words and their multiple senses as sets of synonyms. Many researchers (e.g., [12]) have exploited WordNet as a knowledge source to support text analysis tasks. The way we use WordNet in this work is unique in two ways. First, only unstructured and non-disambiguated lexical information - the list of words for all definitions - is exploited to emulate brain knowledge related to words in the text; and second, the information from WordNet is not used to semantically augment a bag-of-words representation but is rather available directly as candidate topic labels. More precisely, given a word extracted sequentially from the text, the algorithm retrieves the set of synonyms for nouns, as well as a short definition and a sentence showing a sample usage. This is done for all senses of a word that can be interpreted as a noun. There is no attempt at selecting the correct sense as is common in traditional natural language processing with word sense disambiguation techniques [28]. Convergence to meaning and disambiguation are by-products of activation and decay. This is an important attribute of the ReAD principles.

Words retrieved from WordNet are called items to distinguish them from words in the text. Items are filtered to remove words that are deemed useless or too general to precisely identify topics (e.g., thing, entity, time, etc). The list of knowledge items $T_j=\{t_1, t_2, \ldots, t_n\}$ contains items retrieved for word w_j, that is, the non-unique words found in all WordNet senses. Knowledge items emulate the specific regions of the brain that are activated when a word is read. An association between each word w_j and its list of knowledge items T_j is kept in the words table W (step 3.2.4), whereas the various parameters related to the computation of activation of each item are kept in the topics table T (step 3.2.6).

The activation is computed based on a modified version of TFIDF [10]. TFIDF is a common measure of word importance in information retrieval, but we use it here to measure the importance of items related to words within a single document. There is therefore no corpus statistics involved, only word related items statistics within an individual document. The activation α_i of item t_i is the product $\alpha_i = tf_i \times idf_i$ where:

$tf_i = q_i / Q$ (originally in information retrieval, *tf* denotes term frequency, but here it is item frequency) q_i is the number of times item i is retrieved from the knowledge source, and Q is the total number of items retrieved that are not stop words, for all words in the current document.

$idf_i = \log(V/v_i)$ (inverse document frequency) V is the total number of words that are not stop words in the current document and v_i is the number of words that trigger retrieval of item i from the knowledge source.

In the algorithm of Fig 1, the values of Q, V, q_i and v_i used in the calculation of activation are updated in the sub-steps of 3.2.6. The computation of activation is delayed until all items for a word have been seen (at step 3.3.3.1.1) because an item may occur more than once in the set retrieved for a given word and thus may have its q_j value incremented repeatedly. It would therefore be pointless to calculate activation before having collected all counts for a word. The incremental establishment of Q, V, q_i and v_i allows for online processing of words within a document. There are other ways to calculate activation, such as simply counting each item occurrence. We used a TFIDF-inspired approach because it appeared to be a judicious choice for the information retrieval related task of topic identification. We will investigate alternatives in future work.

```
1. inputs: text, knowledge source, list of stop words, max retention
time τmax, decay rate γ and number of topics M.
2.  V=0, Q=0, clear items table T and words table W
3.  while there are words in the text:
3.1   get next word wj
3.2   if wj is not in the stop list:
3.2.1   V++
3.2.2   retrieve information about wj from the knowledge
          source: a list of non unique items Tj={t1,t2,…, tn}
3.2.3   remove stop words from Tj
3.2.4   if wj is not in W: store (wj, Tj, τj = τmax) in words
          table W
3.2.5   else: reset τj = τmax for word wj in words table W
3.2.6   for each ti in Tj
3.2.6.1     Q++
3.2.6.2     if ti is already in T: qi ++  and for first occurrence
              of ti in Tj: vi ++
3.2.6.3     else : qi =1, vi =1 and store (ti, qi, vi) in T

3.3   for each word k in W
3.3.1    if  wk != wj  (not the word just read)
3.3.1.1      τk --
3.3.2    Tk = list of items associated with wk in W
3.3.3    for each item i in list Tk
3.3.3.1      if τk > 0
3.3.3.1.1           αi = αi + ( qi / Q * log(V/vi) )
3.3.3.2      else
3.3.3.2.1           αi = αi - αi / (τmax * γ)
4.  All words have been processed: sort the items in T and output the M
items with highest activation
```

Fig. 1 – An algorithm based on the ReAD principles.

Words extracted from the document are stored in a table emulating human short-term memory (STM), which is table W in the algorithm (step 3.2.4). A fundamental idea is that STM has a limited capacity, as is the case with humans [29]. Because of limited STM capacity, a word previously read will eventually be pushed out of STM when a new word is extracted from the text. STM can be interpreted as a sliding window considering a few words of the text at a time. The size of the window can be pre-determined or computed. In the algorithm of Fig 1, STM capacity is defined by the maximum retention time parameter, τ_{max}. When a new word is read from the document, retention time for that word is set to this maximal value. The τ_k parameter associated with each word k (or τ_j associated with word j depending on which algorithm loop we are in: 3.2 for initialization or 3.3 for updates) is used to determine when a word is being pushed out of STM. Each time a new word is read, the window slides to the right to include the next word while the : oldest word is expulsed from STM. The size of the window determines the persistence of un-decaying activation for items associated with a specific word. Once a word looses the focus of attention granted by the window, that is, when it is pushed out of STM, the activation of items associated with the word starts to decay (step 3.3.3.2.1). There are other decay formulae possible. At this point we have only evaluated the linear decay shown in the algorithm.

The role of decay is to help distinguish between relevant and non-relevant items. Activation can be seen as a process of amplification of relevant items while decay

plays the role of a filter to eliminate semantically unimportant items. γ is the decay rate, a larger value meaning a slower decay. Although an item may be undergoing decay because its associated words have lost the focus of attention, if another word is read that activates this same item, the activation will also accumulate. As a direct consequence, in step 3.3, one can observe that an item's activation changes as many times as there are words associated with that item. Besides, an item's activation may be increased due to one word still being in STM and decreased due to another word because it is not (based on the value of the τ_k parameter associated with each word *k*).

The last words of the documents cannot decay entirely and are therefore advantaged. The solution we have implemented in the current implementation is to ignore any item fully activated due to the presence of a word in the last window before the end of the document, but other options are also possible and need to be investigated. This action is omitted from the algorithm of Fig. 1. There may be other variants to the algorithm presented. For example, one might be to allow for cascading activations, where retrieved items recursively propagate activation to other items in a way that might be expected in neural networks. As well, the selection of items for output could be modified in various ways, such as for instance with an activation threshold instead of selecting the *M* most activated ones. These will be tested in future work. Finally, one might argue that the algorithm could be simplified by eliminating everything related to STM and decay. At this point we have not tested this alternative. However, we must point out that such an alternative implementation does not correspond to the inspiration of the human model of reading as embodies in the ReAD principles, in which neural regions activation is maintained through their association with words in STM and decay occurs over time.

5 Empirical Evaluation and Discussion

To evaluate the quality of the topics produced by our algorithm, we conducted two experiments. First, we processed the documents in the Reuter benchmark collection [30] with our implementation of the ReAD principles, with clustering algorithms and using keywords extracted directly from the documents with TFIDF. Only one topic per document was retained (i.e. *M*=1 in the algorithm) with ReAD. We tested two text clustering algorithms, k-means as a baseline and state-of-the-art spherical k-means [32]. The quality of the results obtained was evaluated with F1, which is a common quality measure in text classification and clustering. F1 results are in the range [0, 1] with 1 being the best quality. Due to space limitation, we refer the reader to [31] for details on the F1 quality metric and its use. Results are shown in table 1, with ReAD having achieved the best F1 quality.

Table 1. Results

Technique	**F1**
ReAD	0.33
Spherical k-means clustering	0.29
K-Means clustering	0.23
TFIDF	0.16

For the second experiment, we turned to human assessments. This answers the important question of whether the topics generated by the program are actually intelligible and evocative of the document content to a human user. According to the independent human assessors, 36% of the topics found by the ReAD algorithm were judged to be acceptable to perfect.

The human assessment scores supplement the F1 quality evaluations, confirming that the algorithmic implementation of the ReAD principles can establish the semantic content of text documents. The level of success is still relatively low but it is comparable to clustering techniques. It is important to note that WordNet is an imperfect replacement for the richness of knowledge found in an actual human brain, as specified in the ReAD principles. Using the limited form of knowledge present in WordNet has the advantage of demonstrating a baseline of what can be achieved, so one can imagine the possibilities with better, richer knowledge such as what can be found on the Web. Nevertheless, it appears quite an accomplishment to obtain a correct identification of topics in 36% of the cases merely with activation and decay of items obtained from a general-purpose lexical source like Wordnet. There is, after all, no special purpose knowledge handcrafting and no traditional semantic or even syntactic analysis. This is also achieved without a large corpus for training, so that each new document can be processed independently and novelty can thus be handled accordingly. Hence, the ReAD algorithm provides a major advantage over existing techniques, namely autonomy and adaptability. Moreover, there are a variety of improvements to be explored, the technique introduced here being in its infancy. Notably, about 30% of words were not found in WordNet and no attempt has been made to exploit words other than nouns. An interesting question that thus needs to be looked into is the effect of using information on all words. For instance, a more complete source of knowledge such as the World Wide Web or Wikipedia could be exploited, as others have done with other techniques [33, 34]. Hence, the web itself could potentially be exploited to make sense of itself. This would be an exciting endeavour in the context of the Semantic Web [35]. We are currently conducting more comprehensive evaluations, examining different variants and parameterization of the algorithm, comparing with other techniques and performing more user assessments.

6 Conclusions

We presented an autonomous and adaptable approach that eliminates the problem of ambiguity and captures the general meaning of a text document. The fundamental ReAD principles behind the approach are: first, the retrieval and activation, for each word read sequentially from the text, of a set of items – simply other words – from a general-purpose, domain independent knowledge source. Second, the decay of infrequent items activation and incremental augmentation for those that occur repeatedly. Third, the convergence over time, as words are read, onto a few discriminately activated items that represent the main concepts discussed in the text, or in other words, its topics.

The principles are different from existing text mining techniques and are inspired by the way humans may be processing information when they read. The innovative nature of the principles avoids computationally complex NLP and the issue of lack of autonomy due to human intervention. As well, it eliminates the dependence of large text corpora, allowing for the processing of single text in isolation and offering adaptive processing in the face of novelty. The algorithm performs computational determination of the general semantic nature of text – its thematic or conceptual content, or in other words, its topics - with general-purpose lexical knowledge. The approach abandons the standard vector and bag-of-word representation, rather harnessing the order and interdependence of words to compute meaning, and this without conventional syntactic and semantic analysis, without task specific knowledge acquisition, and without training.

References

1. Blei, D., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation, Journal of Machine Learning Research (3), 993-1022 (2003).
2. Landauer, T. K., Dumais, S. T. : Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review 104 (2), 211-240 (1997).
3. McNamara, D.S.: Computational methods to extract meaning from text and advance theories of human cognition, Topics in Cognitive Science 3(1), 3--17 (2011).
4. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (2002).
5. Qi, X., Davison, B. D.: Web page classification: Features and algorithms. ACM Comput. Surv. 41, 2 (2009).
6. Jain, A. K., Murty, M. N., Flynn, P. J.: Data clustering: a review. ACM Comput. Surv. 31, 3 (1999).
7. Feldman, R. Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data (Cambridge University Press, NY, 2006).
8. Massey, L.: On the quality of ART1 text clustering. Neural Networks, 16, 5-6 (2003).
9. Salton, G., Lesk, M.E.: Computer evaluation of indexing and text processing. J. ACM, 15, 1 (1968).
10. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. of Doc. 28 , 1 (1972).
11. Halliday, M. A. K. , Hasan, R.: Cohesion in English (Longman Pub Group, NY, 1976).
12. Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In Proceedings of Semantic Web Workshop, the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (ACM, NY, 2003).
13. Hu, J., Fang, L., Cao, Y., Zeng, H. , Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 179-186 (ACM, NY, 2008).
14. Scott, S., Matwin, S.: Feature engineering for text classification. In Proceedings of 16th International Conference on Machine Learning, 379-388 (1999).
15. Lenat, D. B.: CYC: A Large-Scale Investment in Knowledge Infrastructure. Commun. ACM 38, 11 (1995).

16. Milne D., Witten I.H.: Learning to link with wikipedia. In Proceeding of the 17th ACM conference on Information and knowledge management (CIKM '08). ACM, New York, NY, USA, 509-518. (2008).
17. Kim H.L., Scerri S., Breslin J.G., Decker S., Kim H.G.: The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications (DCMI '08). Dublin Core Metadata Initiative 128-137 (2008).
18. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval, 2(4), 303-336 (2000).
19. Velardi, P., Navigli, R., D'Amadio, P.: Mining the Web to Create Specialized Glossaries, IEEE Intelligent Systems, 23(5), 18-25, (2008).
20. Wong, W., Liu, W., Bennamoun, M.: A probabilistic framework for automatic term recognition. Intelligent Data Analysis 13(4), 499-539 (2009).
21. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1(4):390 (1957).
22. Cabre-Castellvi, T., Estopa, R. Vivaldi-Palatresi, J.: Automatic term detection: A review of current systems. In D. Bourigault, C. Jacquemin, and M.C. L'Homme, Eds, Recent Advances in Computational Terminology (John Benjamins, 2001).
23. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing - Volume 10, Association for Computational Linguistics, Morristown, NJ, 216-223 (2003).
24. Milne, D. N., Witten, I. H., & Nichols, D. M.: A knowledge-based search engine powered by wikipedia. In Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management, 445-454 (2007).
25. Jarvella, R. J.: Syntactic processing of connected speech. J. Verb. Learn. Verb. Behav. 10 (1971).
26. Just, M.A., Carpenter P.A.: A capacity theory of comprehension: Individual differences in working memory. Psychol. Rev. 99 (1992).
27. Fellbaum, C.: WordNet: An Electronic Lexical Database (1998).
28. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2 (2009).
29. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63 (1956).
30. Lewis, D.D. : Reuters-21578 Distribution 1.0. Last retrieved 22 April 2010, http://www.daviddlewis.com/resources/testcollections/reuters21578
31. Massey L.: Evaluating and Comparing Text Clustering Results. In Proceedings of 2005 IASTED International Conference on Computational Intelligence (2005).
32. Dhillon, I.S., Modha, D.M.: Concept Decompositions for Large Sparse Text Data using Clustering. Mach. Learn. 42, 1 (2001).
33. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In Proceedings of the 21st National Conference on Artificial intelligence, 1301-1306 (2006).
34. Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., Zhang, T.: Classifying search queries using the Web as a source of knowledge. ACM Trans. Web 3, 2 (2009).
35. Berners-Lee, T., Hendler, J. Lassila, O.: The Semantic Web. Sci. Am. 284, 5 (2001).