# Effective Extraction of Thematically Grouped Key Terms From Text

**Maria Grineva** **Maxim Grinev** and **Dmitry Lizorkin**

Institute for System Programming of Russian Academy of Sciences
25 A. Solzhenitsyna
Moscow, Russia 109004

### Abstract

We present a novel method for extraction of key terms from text documents. The important and novel feature of our method is that it produces groups of key terms, while each group contains key terms semantically related to one of the main themes of the document. Our method bases on a combination of the following two techniques: *Wikipedia-based semantic relatedness measure of terms* and *algorithm for detecting community structure of a network*. One of the advantages of our method is that it does not require any training, as it works upon the Wikipedia knowledge base.

Our experimental evaluation using human judgments shows that our method produces key terms with high precision and recall.

## Introduction

*Key terms* (sometimes referred to as keywords or key phrases) are set of significant terms in a text document that give high-level description of the content for readers. Key terms extraction is a basic step for various tasks of natural language processing, such as document classification, document clustering, text summarization and inferring a more general topic of a text document (Manning and Schtze 1999). In this paper we propose a method for key terms extraction from text using Wikipedia as a rich resource of terms semantic relationships.

Wikipedia (www.wikipedia.org) is a free online encyclopedia that has grown to become one of the largest online repositories of encyclopedic knowledge. It contains millions of articles available for a large number of languages. As for September 2008, English Wikipedia contains over 2.5 million articles (over 6 million if consider redirects). With its extensive network of cross-references, categories, portals, redirect and disambiguation pages it has become an exceptionally powerful resource for this work and many other natural language processing (NLP) and information retrieval (IR) applications.

The background of our method consists in the following two techniques: measure of terms *semantic relatedness* computed over Wikipedia corpus; and network analysis algorithm, namely, *Girvan-Newman algorithm* for detecting community structure in networks. We give a short overview of these techniques below.

Identifying the semantic relatedness of concepts within Wikipedia seems to be a natural way to build the usable tool from Wikipedia knowledge for NLP/IR applications. During the last three years, there have appeared a good few of works on computing Wikipedia-based semantic relatedness using different methods (Milne and Witten 2008; Milne 2007; Gabrilovich and Markovitch 2007; Strube and Ponzetto 2006; Turdakov and Velikhov 2008). See (Milne 2007) for an insightful comparison of the many of existing Wikipedia-based semantic relatedness measures. While our method does not provide any requirements on the way of computing semantic relatedness, the effectiveness of our method depends on the effectiveness of the exploited semantic relatedness measure. For the evaluation of our method in this paper we used semantic relatedness measure described in (Turdakov and Velikhov 2008).

Having semantic relatedness measure of terms allows us to build *semantic graph* for all the terms of a processed text document. Semantic graph is a weighted graph where *vertices* are terms, *edge* between a pair of terms means that these two terms are semantically related, the *weight* of the edge is the semantic relatedness measure of the two terms. We noticed that the graph constructed in this way has an important feature: the terms related to the common topics bunch up into dense subgraphs or *communities*, and the most massive and densely interconnected groups typically correspond to the main topics of the processed document! The novelty of our method consists in applying network analysis algorithm for detecting community structure of the constructed semantic graph, and then selecting the densest groups of terms that inherently represent thematically grouped key terms of the document.

The discovery and analysis of community structure in networks - natural divisions of network nodes into densely connected subgroups - has been well studied. Many algorithms have been proposed and applied with great success to social networks (Wasserman, Faust, and Iacobucci 1994), citation networks (Redner 1998; de Solla Price 1965), purchasing network (Clauset, Newman, and Moore 2004), biochemical networks (Kauffman 1969) and many others. However, to the best of our knowledge, there are no applications of community detection algorithms to the Wikipedia-based net-

works. In our method we use the algorithm invented by M. Newman and M. Girvan (Newman and Girvan 2004) that has been proved to be highly effective at discovering community structure in both computer-generated and real-world network data.

The rest of the paper is organized as follows. The next section reviews related work and positions our method among existing methods. This is followed by a description of our method, its evaluation against manually defined ground-truth and conclusion.

## Related Work

There are classical approaches for extraction of key terms in statistical natural language processing: *tf.idf* and *collocation analysis* (Manning and Schtze 1999). The tf.idf (term frequency-inverse document frequency) is a metric often used in information retrieval and text mining (Salton and Buckley 1988). This metric is a statistical measure used to evaluate how important a term is to a document in a collection or corpus. The importance is proportional to the number of times a term appears in the document divided by the frequency of the word in the corpus. While tf.idf can be used to extract a single-word key terms, collocation analysis is used to identify key phrases. *Chi-square independence test* (Manning and Schtze 1999) is often used for collocation analysis: one can determine if a sequence of terms (phrase) co-occur more often than would be expected by chance. Supplementing tf.idf with collocation analysis allows extracting and ranking candidate key phrases of the document. The two approaches require some document collection to gather terms statistics from or a *training set*. The quality of result (extracted key terms) in these approaches depends on the quality of training set. The great advantage of these approaches is their simplicity of implementation and satisfactory quality of results when training set successfully fits the application data, and consequently these approaches are widely used in many practical applications. An interesting fact that we would like to note here is that there are works (for example, (Mihalcea and Csomai 2007; Mihalcea 2007; Dakka and Ipeirotis 2008; Medelyan, Witten, and Milne 2008)) showing that Wikipedia can serve as a good training set.

There is an alternative class of approaches to solve many NLP tasks (key terms extraction being only one of them) that base on using measures of terms semantic relatedness (Budanitsky and Budanitsky 1999), and our work belongs to this class of approaches. Terms semantic relatedness can be inferred from a dictionary or thesaurus (for example, WordNet (Miller et al. )), but here we are interested in terms semantic relatedness derived from Wikipedia. Wikipedia-based semantic relatedness measure for two terms can be computed using either the links found within their corresponding Wikipedia articles (Milne and Witten 2008; Milne 2007), or the article's textual content (Gabrilovich and Markovitch 2007). There is a bunch of works on using Wikipedia-based semantic relatedness to solve the following basic NLP/IR tasks: word sense disambiguation (Mihalcea 2005; Sinha and Mihalcea 2007; Medelyan, Witten, and Milne 2008; Turdakov and Velikhov 2008), topic inferring (Syed, Finin, and Joshi 2008), classification (Janik and Kochut 2008), coreference resolution (Strube and Ponzetto 2006). Though, to the best of our knowledge, there are no works on extracting key terms using Wikipedia-based terms semantic relatedness, the work (Janik and Kochut 2008) is the closest to ours. In (Janik and Kochut 2008) the idea to apply graph analysis techniques is presented in its rudimentary form: the most central terms in a semantic graph are identified using betweenness centrality measure (Newman and Girvan 2004). These terms further constitute a basis for document categorization.

We point out the following advantages of our method:

- Our method does not require any training as opposed to the described traditional approaches. With its scale and constant maintenance, Wikipedia covers a huge number of different domains and remains up-to-date. Thus, almost any document, most part of which terms are described in Wikipedia, can be processed by our method.
- Key terms are grouped by document themes, and method provides as many groups as there are different themes covered in the document. Thematically grouped key terms can significantly improve further inferring of document topics (using, for example, spreading activation over Wikipedia categories graph as described in (Syed, Finin, and Joshi 2008)) and categorization (Janik and Kochut 2008).
- Our method is highly-effective from the quality viewpoint. Our evaluation using human judgments shows that it produces key terms with high precision and recall.

## Method for Key Terms Extraction

The method consists of the five steps that we describe in detail in the following subsections: (1) candidate terms extraction; (2) word sense disambiguation; (3) building semantic graph; (4) discovering community structure of the semantic graph; and (5) selecting valuable communities.

### Candidate Terms Extraction

The goal of this step is to extract all terms from the document and for each term prepare a set of Wikipedia article that can describe its meaning.

We parse the input document and extract all possible n-grams. For each n-gram we construct its variations using different morphological forms of its words. We search for all n-gram's variations among Wikipedia article titles. Thus, for each n-gram a set of Wikipedia articles can be provided.

Constructing different morphological forms of words allows us not to miss a good fraction of terms. For instance, *"drinks"*, *"drinking"*, and *"drink"* can be linked to the two Wikipedia articles: *"Drink"* and *"Drinking"*.

### Word Sense Disambiguation

At this step we need to choose the most appropriate Wikipedia article from the set of candidate articles for each ambiguous term extracted on the previous step.

It is an often situation in natural language when a word is *ambiguous*, i.e. carries more than one meaning, for example:

the word "platform" can be used in the expression "railway platform", or it can refer to hardware architecture or a software platform. The correct sense of an ambiguous word can be selected based on the context where it occurs, and correspondingly the problem of *word sense disambiguation* is defined as a task of automatically assigning the most appropriate meaning (in our case, the most appropriate Wikipedia article) to a word within a given context.

There is a number of works on disambiguating terms using Wikipedia (Turdakov and Velikhov 2008; Medelyan, Witten, and Milne 2008; Sinha and Mihalcea 2007; Mihalcea 2005; 2007). For evaluation in this paper we used method described in (Turdakov and Velikhov 2008). In (Turdakov and Velikhov 2008) authors make use of Wikipedia's disambiguation and redirect articles to obtain candidate meanings of ambiguous terms. For each ambiguous term disambiguation page contains all of the term's meanings, which are separate articles in Wikipedia with their own link structure. For example the article "platform (disambiguation)" contains 17 meanings of the word "platform". Then in (Turdakov and Velikhov 2008) semantic relatedness measure is used to pick the meaning that has the highest relevance to the context where the ambiguous term appeared.

It is a typical problem with traditional key terms extraction techniques when nonsense phrases such as e.g. "using", "electric cars are" appear in the result. Using Wikipedia articles titles as a controlled vocabulary, allows us to avoid this problem, all of the key terms produced by our method are acceptable phrases.

Result of this step is a list of terms, where each term is assigned with a single Wikipedia article that describes its meaning.

### Building Semantic Graph

At this step we build a semantic graph from a list of terms obtained on the previous step.

Semantic graph is a weighted graph where each *vertex* is a term, *edge* between a pair of vertices means that the two terms corresponding to these vertices are semantically related, the *weight* of the edge is the semantic relatedness measure of the two terms.

Figure 1 shows semantic graph built from a news article *"Apple to Make ITunes More Accessible For the Blind"*. This article tells that the Massachusetts attorney general's office and the National Federation of the Blind reached an agreement with Apple Inc. under which it will make its music download service (ITunes) accessible to the blind consumers using screen-reading software. In Figure 1 you can see that terms related to *Apple Inc.* and *Blindness* constitute two dominant communities, and terms like *Student*, *Retailing* or *Year* become peripheral and weakly connected or not connected at all to other terms.

An important observation is that disambiguation mistakes tend to become isolated vertices in a semantic graph and not to adjoin to dominant communities.

### Discovering Community Structure of the Semantic Graph

At this step we discover community structure of the semantic graph built on the previous step. We use Girvan-Newman algorithm for this purpose (Newman and Girvan 2004). The algorithm divides the input graph into a number of subgraphs that are likely to be dense communities.

To estimate the goodness of the certain graph partition, authors of (Newman and Girvan 2004) propose the notion of *modularity*. Modularity (Newman and Girvan 2004) is a property of a network and a specific proposed division of that network into communities. It measures when the division is a good one, in the sense that there are many edges within communities and only a few between them. In practice, modularity values that fall in the range from about 0.3 to 0.7 indicate that network has quite a distinguishable community structure.

We observed that semantic graphs constructed from an average text document (one page news article, or a typical academic paper) have modularity values between 0.3 and 0.5.

### Selecting Valuable Communities

At this step we need to rank term communities in a way that highest ranked communities would contain terms semantically related to the main topics of the document (key terms), and the lowest ranked communities contain not important terms, and possible disambiguation mistakes (terms which meaning was chosen wrong on the second step).

Ranking is based on the *density* and *informativeness* of communities. Density of a community is a sum of weights of all inner-community edges divided by the number of vertices in this community.

While experimenting with traditional approaches, we observed that using tf.idf measure of terms can help ranking communities in a proper way. tf.idf measure gives higher values to the named entities (for example, *Apple Inc.*, *Steve Jobs*, *Braille*) than to general terms (*Consumer*, *Year*, *Student*). We compute tf.idf measure of terms using Wikipedia corpus as described in (Medelyan, Witten, and Milne 2008). *Informativeness* of a community is a sum of tf.idf measure of all terms in a community divided by the number of terms.

Eventually, the rank value assigned to each community is its density multiplied by its informativeness, and communities are then sorted according to this value.

Application that uses our method to extract key terms is free to take any number of the top-ranked communities, however our recommendation is 1-3 top-ranked communities.

## Evaluation

In this section we discuss the experimental evaluation of our techniques. Since there is no standard benchmark for evaluating the quality of the extracted key terms, we conducted a human study, trying to evaluate the "precision" and "recall" of the extracted key terms. We choose 30 blog posts from the following technical blogs: *Geeking with Greg* by Greg Linden, *DBMS2* by Curt Monash and *Stanford Infoblog* by people from Stanford Infolab. Five persons from

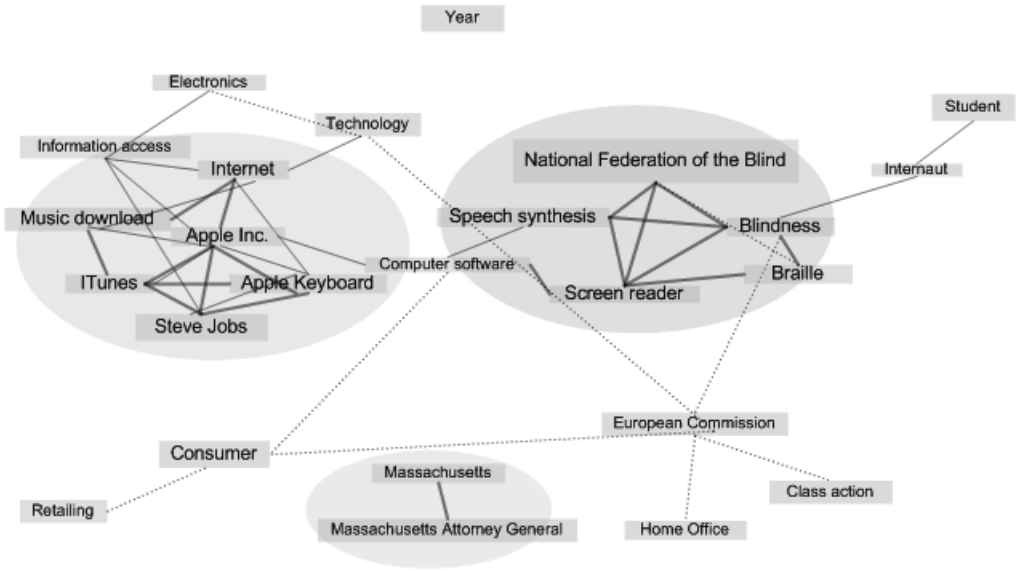


Figure 1: Semantic graph built from the news article *"Apple to Make ITunes More Accessible For the Blind"*

the MODIS department of the Institute for System Programming took part in the evaluation. Each person was asked, for each blog post, to read it and to identify from 5 to 10 of its key terms. Each key term must present in the blog post, and must be identified using Wikipedia article names as the allowed vocabulary. We instructed our evaluation participants to choose key terms that cover several main topics of the blog post. Eventually, for each blog post we considered a key term to be valid if at least two of the participants identified the same key term from this blog post. We considered Wikipedia redirection article names and the article which they redirect to being "the same" term, as they inherently represent synonyms.

The techniques presented in the paper were implemented with the following architectural design principles. For achieving the best computational efficiency, all necessary data concerning Wikipedia articles names, Wikipedia link structure and statistical information about the Wikipedia corpus (for instance, terms tf.idf) are kept in main memory. With the recent Wikipedia being quite a large knowledge base, our Wikipedia knowledge base takes 4.5 Gigabytes in the main memory. A dedicated machine with 8Gb RAM was used for the evaluation, and client applications access the knowledge base via remote method invocation. With the requirement for having access to similarity scores for virtually *every* term pair in Wikipedia, similarity scores are not computed offline in advance, but are rather computed on demand on the fly using Wikipedia link structure (Turdakov and Velikhov 2008).

## Recall

We define recall as the fraction of the manually extracted key terms that were also extracted by our method:

$$\frac{|\{\text{manually extracted}\} \bigcap \{\text{automatically extracted}\}|}{|\{\text{manually extracted}\}|},$$

with $\{$manually extracted$\}$ denoting the set of all terms identified for a document by humans, $\{$automatically extracted$\}$ denoting the set of all terms extracted by the suggested technique for the same document and $|S|$ denoting the number of items in a set $S$.

For the 30 blog posts we have got 180 key terms extracted manually, 297 key terms were extracted by our method, 123 of manually extracted key terms were also extracted by our method. Thus, we have got the recall equals to **68%**.

## Precision

First, we estimate precision using the same methodology that we used for estimating recall: we define precision as the fraction of the terms automatically extracted by our method that were also extracted by humans:

$$\frac{|\{\text{manually extracted}\} \bigcap \{\text{automatically extracted}\}|}{|\{\text{automatically extracted}\}|}.$$

According to our test collection the precision of our method equals to **41%**.

## Revision of Precision and Recall

However, we have revisited the measuring of precision and recall according to the specifics of our method. The important thing is that our method, on average, extracts more terms than a human. More precisely, our method typically extracts more related terms in each thematic group than a human. For example, consider Figure 1, for the topic related to *Apple Inc.* our method extracts terms: *Internet*, *Information access*, *Music download*, *Apple Inc.*, *ITunes*, *Apple Keyboard*, *Steve Jobs*; while a human typically identifies less, and tends to identify named entities: *Music download*, *Apple Inc.*, *ITunes* and *Steve Jobs*. That means that, possibly, sometimes our method produces better terms coverage for a specific topic than an average human. And this is a reason that we measure the precision and recall in another way also.

Each evaluation participant was asked to revisit his key terms in the following way. For each blog post he was provided with key terms extracted automatically for this blog post. He had to review these automatically extracted key terms and, if possible, extend his manually identified key terms with some from the automatically extracted set. It appeared that humans indeed found out relevant key terms that they had not extracted before, and extended their key terms.

After this revision we have got 213 manually extracted key terms for the 30 blog posts (instead of 180), thus, evaluation participants had added 33 new key terms from the automatically extracted set. The recall is then equals to **73%** and the precision is **52%**.

### Computational Efficiency

When experimenting with the implementation, we observed that most computation time was consumed by (i) text parser for extracting candidate terms from input document and (ii) semantic graph construction that is essentially obtaining similarity scores for candidate term pairs. Compared to these preliminary steps, the running time for the remaining steps of the algorithms is negligible, with both community discovery and selection of valuable communities being essentially linear (Clauset, Newman, and Moore 2004) in the number of edges in the semantic graph.

On average, it takes about 4 minutes to extract key terms from 100 blog posts.

## Conclusion

We presented a novel method for extracting key terms from a text document. One of the advantages of our method is that it does not require any training, as it works upon the Wikipedia-based knowledge base. The important and novel feature of our method is that it produces groups of key terms, while each group contains key terms related to one of the main themes of the document. Thus, our method implicitly identifies main document themes, and further categorization and clustering of this document can greatly benefit from that.

Our experimental results, validated by human subjects, indicate that our method produces high-quality key terms comparable to the ones by state-of-the-art systems developed in the area. Evaluation showed that our method produces key terms with 73% recall and 52% precision, that we consider being significantly high.

We are inspired by (Dakka and Ipeirotis 2008) in conducting a more extensive evaluation using Amazon Mechanical Turk service.[1] This service offers access to a community of human subjects and proposes tools to distribute small tasks that require human intelligence.

We observed that our method is good for cleaning up the content of the document from non-important information and possible disambiguation mistakes. That means that it can have a great potential for extracting keywords from Web pages which usually flooded with menus, button titles and advertisements. We plan to investigate this side of the method further.

[1] www.mturk.com

## References


Budanitsky, A., and Budanitsky, A. 1999. Lexical semantic relatedness and its application in natural language processing. Technical report, University of Toronto.

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E* 70:066111.

Dakka, W., and Ipeirotis, P. G. 2008. Automatic extraction of useful facet hierarchies from text databases. In *ICDE*, 466–475. IEEE.

de Solla Price, D. J. 1965. Networks of scientific papers. *Science* 169:510–515.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, 1606–1611.

Janik, M., and Kochut, K. J. 2008. Wikipedia in action: Ontological knowledge in text categorization. *International Conference on Semantic Computing* 0:268–275.

Kauffman, S. A. 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* 22(3):437–467.

Manning, C. D., and Schtze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Medelyan, O.; Witten, I. H.; and Milne, D. 2008. Topic indexing with wikipedia. In *Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08)*.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 233–242. New York, NY, USA: ACM.

Mihalcea, R. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 411–418. Morristown, NJ, USA: Association for Computational Linguistics.

Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation.

Miller, G. A.; Fellbaum, C.; Tengi, R.; Wakefield, P.; Langone, H.; and Haskell, B. R. Wordnet: a lexical database for the english language. http://wordnet.princeton.edu/.

Milne, D., and Witten, I. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08)*.

Milne, D. 2007. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC)*.

Newman, M. E. J., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69:026113.

Redner, S. 1998. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B* 4:131.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

Sinha, R., and Mihalcea, R. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, 363–369. Washington, DC, USA: IEEE Computer Society.

Strube, M., and Ponzetto, S. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)*, 1419–1424.

Syed, Z.; Finin, T.; and Joshi, A. 2008. Wikipedia as an Ontology for Describing Documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press.

Turdakov, D., and Velikhov, P. 2008. Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Colloquium on Databases and Information Systems (SYRCoDIS)*.

Wasserman, S.; Faust, K.; and Iacobucci, D. 1994. *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.