Recognition of Semantically Related Articles in Wikipedia

Ivan Valenčík[*]
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
valencik@gmail.sk

Abstract

Wikipedia has become the most used encyclopaedia in the world. The English edition itself has around 4 million articles and there are together more than 280 language editions of Wikipedia. However there are big differences in their scope and detail. This difference is an opportunity to improve the content of the less detailed Wikipedias by information extracted from the big ones. In this paper we propose a method for recognition of semantically related articles based on connections between Wikipedia editions. We evaluate an implementation of this method using a MapReduce programming model and a HBase database system.

Categories and Subject Descriptors

E.1 [**Data**]: Data structures—*graphs and networks*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, search process, selection process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*distributed systems*

Keywords

recognition, semantic relatedness, Hadoop, HBase

1. Introduction

Wikipedia is a free, web-based, collaborative, multilingual encyclopaedia project. It was launched in 2001 and today it is the biggest and most widely used encyclopaedia in existence [6]. We can look at every language edition of Wikipedia as a network with nodes corresponding to its articles and links corresponding to the hyperlinks between them. In comparison to classical encyclopaedias Wikipedia grows all the time. Authors are expected to follow a set of editorial and structural rules.

Despite Wikipedia having about 280 language editions, only four editions of Wikipedia have more than one million articles. There are also big differences in their quality [7]. Articles in different Wikipedias about the same concept are written independently, so they are not just translations from other editions. However it can be still argued, that all Wikipedia editions are describing the same knowledge. Each article in Wikipedia should describe a single concept and there should be only one article for each concept. Based on this we could presume, that the network representing Wikipedia should be the same in all language editions and it should differ only in its completeness. Relations between concepts should be consistent as well.

This is not the case in practice, because smaller Wikipedias describe in one article more concepts than there are identified in bigger Wikipedias and even in big Wikipedias authors are often not consistent in the way they write about some topics. This is partly caused by natural evolution of articles but also by differences in editorial rules. Despite these discrepancies it might be possible to enhance smaller Wikipedias, but also big ones, solely by comparing network structure of language editions. Small Wikipedias do not have only small number of articles, but in general they are also of lower quality. Their articles are in comparison with English edition of Wikipedia shorter, less structured, not linked to all relevant articles and they are often not assigned to proper categories.

It is very hard to work directly with the whole Wikipedia because of its scale. For this reason we use a MapReduce programming model introduced by Google, which enables us to work with a large data set on a cluster of computers without the need to go into details of distributed computing. We store the working data and results in a highly scalable database based on a BigTable database system designed by Google.

The paper is structured as follows. In section 2 we describe Wikipedia and its relevant characteristics. Sec-

[*] Doctoral degree study programme in field Software Engineering. Supervisor: Dr. Peter Lacko, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava.
Work described in this paper was presented at the 8th Student Research Conference in Informatics and Information Technologies IIT.SRC 2012.

Valenčík, I. Recognition of Semantically Related Articles in Wikipedia. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 4, No. 2 (2012) 65-68

tion 3 contains a description of a proposed method. In section 4 we describe the implementation of the method. In section 5 we evaluate a basic version of the method and in section 6 we give conclusion.

2. Data

Wikipedia consists of many kinds of pages [6]. Articles are the basic unit of information in Wikipedia. They are written in form of free text and follow a set of editorial and structural rules, the most important for us being that each article describes a single concept, and there is a single article for each concept. Articles contain links expressing their relation to other articles.

All Wikipedia editions are available for download to interested users[1]. In this paper we use a Slovak edition as a small Wikipedia that is being enriched and a Danish Wikipedia which serves us as a big edition based on which additions are being made. We take into consideration only simple articles and hyperlinks between them.

3. Method

A purpose of the proposed method is to recognize semantically related articles of a selected article using interlingual links between Wikipedia editions and a network structure of all concerned language editions. We use a premise that semantically related articles are in the Wikipedia network direct neighbours much more often than unrelated articles or articles that are related only in a given context.

The interlingual links are in general correct; however a certain number of links is imprecise or wrong. Their removal can be formalized as an optimization task based on graph repair operations as in [3]. This still does not provide certainty that all incorrect interlingual links are removed nor does it solve the problem of missing links. The incorrect links are only a small subset of all links; therefore we choose to ignore them. We also do not try to interconnect articles which have no interlingual connection to the other used language editions at all. In case of articles that have the interlingual link at least in one direction we are able to add the link in the opposite direction with a sufficient confidence that we are not adding more mistakes.

The recognition itself can be divided into 2 phases. In the first phase we recognize candidate articles for addition. In the simple form of the method we enrich one language edition by information from the second edition. To do this we use a following algorithm:

1. We pick an article a_1 in Wikipedia W_1 which is to be enriched.

2. Using its interlingual link we find a_2 which is a counterpart of a_1 in W_2.

3. We find articles $A_2 = \{a_{21}, a_{22}, a_{23}, \ldots\}$ from W_2 that are in the surrounding of a_2.

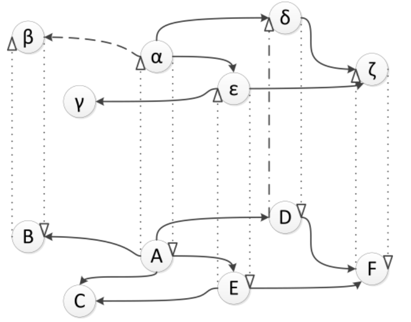
4. Using the interlingual links we find articles $A_1 = \{a_{11}, a_{12}, a_{13}, \ldots\}$ which are counterparts of A_2 articles in W_1.

5. We find articles $A_1' = \{a_{11}, a_{14}, \ldots\}$ from W_1 which have an incoming link from a_1.

6. We recognize articles $R = \{a_{12}, a_{13}\}$ by taking a relative complement of A_1' in A_1 ($R = A_1 \backslash A_1'$).

In this algorithm we enrich the article a_1 in the Wikipedia edition W_1 using articles found in the second Wikipedia edition W_2. The surrounding in the step 3 can by chosen in many ways. In this paper we focus on the simplest approach taking into account only outgoing links, i.e. directed links based on hyperlinks leaving the chosen article. We can extend this method by using also incoming links. This would inevitably cause that we recognize more candidate articles. Alternatively we can use also different approaches, e.g. choosing only some neighbouring articles based on the position of the hyperlink in the text (the first paragraph, infoboxes, etc.) or choosing also articles that are not direct neighbours of the enriched article based on some metric. This method needs to be picked with care, because it is the costliest part of the algorithm as it requires a fetch operation for every considered node.

Figure 1: Example of link recognition.



If we cannot find the interlingual link in the step 2, the whole algorithm fails. In the step 4 we get one less candidate article for each missing interlingual link. See Figure 1 for an example of links that could be added. The dash-dotted link from D to δ is an example of interlingual link that can be fixed. The described algorithm allows us to add a link from α to β. We are not able to recognize the connection between α and γ because of the missing interlingual link. We would need more complex method to evaluate a relation of α to ζ.

In the second phase of the recognition we order the candidate articles and select the desired number of them as a final result. The ordering can be based on various metrics. One option is to order the articles by the number of articles referencing them, i.e. the number of incoming links. This method can be further divided based on whether we consider duplicate incoming links to be just one reference and or multiple references. The next viable metric is a local PageRank [2]. We need only a simple version of this algorithm as we do not need to compare results globally, because we use it to make only a local decision. There is therefore no need for elaborate estimation of boundary node values and normalization. Another option is to order the candidate articles by their semantic relatedness with the enriched article. We can achieve this by using Gabrilovich's and Markovitch's Explicit Semantic Analysis devised specifically for Wikipedia [5]. These orderings can be also combined by assigning them a weight according to their importance.

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

Table 1: Evaluated terms with the number of added links and the achieved precision.

Article	Links added	Precision	Article	Links added	Precision
Velvet Revolution	2	100.00%	Cat	4	25.00%
American Civil War	76	22.37%	Mars	25	24.00%
Angela Merkel	17	35.29%	Lilium	6	66.67%
Milan Rastislav Štefánik	4	75.00%	Berlin	65	27.69%
Philosophy	27	96.30%	Google	10	60.00%
Thirty Years' War	41	36.59%	Poprad	1	100.00%
Søren Kierkegaard	27	59.26%	Lynx	9	88.89%
Hans Christian Andersen	21	33.33%	Sun	101	25.74%
Informatics	2	50.00%	Aarhus	35	14.29%
Mathematics	21	95.24%	Oak	3	100.00%

The number of articles selected into the result can be given, it may depend on some metric representing the reliability of selection (e.g. achieved relatedness) or it can be all of the candidate articles.

4. Implementation

We implemented a basic version of the described method using a Hadoop software framework implementing the MapReduce programming model and an HBase database system implementing the BigTable storage system. MapReduce is used for a work with large data sets. Tasks are specified by a *map* function that processes key/value pairs into intermediate key/value pairs and a *reduce* function that merges and processes all intermediate pairs with the same key [4]. A BigTable storage system is a sparse, distributed, persistent multidimensional sorted map indexed by a row key, column key, and a timestamp [1].

The proposed method can be divided into 3 stages described in this section. In each stage we are executing a one type of the Hadoop job. When working with two languages we need to execute the job in the first stage two times and job in the other two stages once.

4.1 Data processing

A purpose of this stage is to transfer all relevant data from the input Wikipedia XML file to the HBase database. Data are processed with a MapReduce job which is run once for every language edition. The nodes from the XML file containing articles are the input of these jobs. We process only simple articles in the *map* function; all other categories of pages are ignored. All articles are given a unique identifier consisting of the language code and the title, e.g. "`sk:Bratislava`". This naming convention ensures that all the articles from one Wikipedia are grouped together in HBase. We also store a list of all outgoing links and chosen language editions. A key/value pair is emitted for every outgoing links in which the target identifier of the link is the key and the value is the processed article identifier. This allows us to build a reverse index in the *reduce* step. We store a list of incoming articles which also allows us to compute how many times the article is referred to.

4.2 Interconnection

A purpose of this stage is to ensure consistency of interlingual links, i.e. make sure that if we have a link in one direction, then there is some connection also in the other direction. It does not necessarily need to be a connection to the same article, because the mapping between language editions is not unambiguous. All articles in the database are used as an input of the *map* function. The function first emits the identifier of the article as a key and a marked list of all interlingual links. Afterwards a key/value pair is emitted for every interlingual link in which the target of the interlingual links is the key and the value is the identifier of the processed article. In the *reduce* function we add all values that are not already present to the marked list. If there are any changes then the list is updated in the database.

4.3 Recognition

In the recognition stage we use only *map* function. All articles from the enriched language edition are used as an input. We apply the method described in section 3. We can use more versions of this method in one MapReduce job to effectively evaluate more approaches. The efficiency gain comes from the fact that the articles are read only once from the database and the ordering method can be used multiple times with one search for the candidate articles. The results must be stored in distinct columns when using more methods at once.

5. Evaluation

We evaluated the method by trying to enrich a Slovak edition of Wikipedia by using the Danish Wikipedia as a big edition. The XML file with the Slovak Wikipedia has 581 991 854 bytes and the Danish Wikipedia has 714 994 390 bytes. In the data processing stage of the Slovak edition the map function processed 2 246 062 key/value pairs and emitted 33 158 098 pairs which were merged into 630 187 pairs to be processed in the reduce function. There were 3 553 609 key/value pairs processed and 47 743 944 pairs emitted which were merged into 773 298 pairs to be processed in the reduce function when processing the Danish Wikipedia. There are 27 893 interlingual links between Slovak and Danish editions of Wikipedia.

In the recognition stage we decided to apply no ordering and evaluate all results. We chose 20 terms from various encyclopaedic topics to evaluate precision of the algorithm. First, we evaluate precision for each of these terms. We do this by deciding how many of the added links are a meaningful addition to the article. We require these links to have some meaning even without the context in which they are used in the bigger edition to be accepted as a usable addition, e.g. we do not accept Virginia as a worthwhile addition to the American Civil War even though the war was fought there. An example of a good addition is Statistics or Integer to the article Mathematics or Battle of Berlin or Prussia to the article Berlin. We also do not take into consideration all links to dates, years, decades, centuries and millennia because they are added very often and without the context they provide very little value and it is easy to reliably algorithmically filter them.

Table 1 depicts evaluated terms with a respective number of added links and an achieved precision. The average precision for the whole evaluation set is 56.5% and there were on average added 24.9 links to every article. Note that almost all linked articles have some connection to the original term; however, our goal is to add only those that can provide useful additional information when reading the article.

6. Conclusions

In this paper we propose a method for recognition of semantically related articles in Wikipedia which uses interlingual links between its language editions. We evaluated this method by using the simple version of the proposed method to enrich the Slovak Wikipedia. The concept proved to be sound even when we used the Danish edition of Wikipedia as the source of new links; which is not considerably more extensive than the Slovak one. We confirmed our assumption that there is so much room for improvement of articles that language editions can benefit even from editions of comparable size. We intend to investigate efficiency of the more complex versions of the method and use bigger and more editions of Wikipedia as the source of the links in future work.

Acknowledgements. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

[1] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: a distributed storage system for structured data. In *Proceedings of the 7th symposium on Operating systems design and implementation*, OSDI '06, pages 205–218, Berkeley, CA, USA, 2006. USENIX Association.

[2] Y.-Y. Chen, Q. Gan, and T. Suel. Local methods for estimating pagerank values. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 381–389, New York, NY, USA, 2004. ACM.

[3] G. de Melo and G. Weikum. Untangling the cross-lingual link structure of wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 844–853, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[5] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India, 2007.

[6] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from wikipedia. *Int. J. Hum.-Comput. Stud.*, 67(9):716–754, Sept. 2009.

[7] Meta-Wiki contributors. List of wikipedias, 2012. [Online; accessed 21-Feb-2012].