Mining relations between Wikipedia Categories

Julian Szymański

Gdańsk University of Technology,
Narutowicza 11/12, 80-952 Gdańsk, Poland,
julian.szymanski@eti.pg.gda.pl

Abstract. The paper concerns the problem of automatic category system creation for a set of documents connected with references. Presented approach has been evaluated on the Polish Wikipedia, where two graphs: the Wikipedia category graph and article graph has been analyzed. The linkages between Wikipedia articles has been used to create a new category graph with weighted edges. We compare the created category graph with the original Wikipedia category graph, testing its quality in terms of coverage.

1 Introduction

Wikipedia is a free encyclopedia available on-line. Its resources are contributed by volunteers and are freely available for edit. Its openness and lack of rigorous and coordinated quality control has been the reason for criticism and controversy, but cooperative editing approach has allowed Wikipedia to increase its content rapidly [1]. The large amount of textual data brings new challenges for algorithms for automatic text processing [2], whose aim is to made knowledge given in the form of text in natural language accessible more easily to the end-users [3]. Wikipeda can also be used as a datasource for text-mining algorithms [4] and deliver very interesting statistical information about language [5]. It also is used as general purpose meta data [6] repository that provide copora for organizing general human knowledge in machine readable form.

In our research studies we are investigating a methods for automatic organization of textual resources. We find Wikipedia to be a very interesting repository where our approaches can be validated. We distinguished three areas of different studies that can help the information in Wikipedia be better accessible:

1. improving existing category system introducing new, significant relations between existing categories,
2. building new categories in the automatic way eg. using text clustering techniques [7],
3. building a new category system based on existing ones using techniques for text classification [8] [9].

In this paper we present the results of the first approach which is the study of building relations between categories used for organization of the documents set.

As data for presented here experiments we used categories available for each article in the Wikipedia and page links connecting articles and categories. Wikipedia categories are less likely to be the target of vandalism, therefore are more reliable than the article data themselves. The problem is that the original Wikipedia system of the categories is made by hand, which makes categories natural, but it causes several problems, one of which being gaps in the relations between categories. Because there is a high number of connections between similar categories omitted and similar articles fall into different categories, the whole system is not coherent which makes it not very useful.

2 Experiment description

Articles in the Wikipedia are connected by page links: any article can link to any number of other articles and vice versa. This connections forms a directed graph – the Article Graph. An article can be assigned to any number of categories, although most articles are assigned only to one. Categories are also interconnected and they form a directed graph – the Category Graph. This graph has been the subject of many studies, and brings a lot of questions: how to organize these categories, how to made connections between articles and categories more reliable, and how to exploit this information as a valid NLP resource [5].

The Article Graph represents relations between encyclopedic entries, while the Category Graph introduces system of the abstract concepts for organizing articles. The categories allow the user to look through the articles on a required, conceptual level. They may allow to find the information on the given subject, the user even does not expect to exists, which is the main advantage in comparison to the traditional approach for searching large repositories of the textual data based on keyword-matching.

In our research we propose a method for adding new information present in the Article Graph to the Category Graph. The Article Graph can be used to compute semantic similarity of a group of the articles [10], and thus form a network of interconnected, general concepts. In our approach we focus on adding new links into existing categories, however an automatic categories construction is also possible [11].

As a result the application of our method a new Category Graph (called Generated Category Graph) with directed and weighted edges is introduced. The nodes (categories) in the new graph remain the same as in the original Category Graph but the new edges are computed from the links present in the Article Graph according to formula R:

$$R(C1, C2, w * n)$$

meaning that there are w articles in the category $C1$, that link to articles in the category $C2$, n is used for weight normalization and is calculated as:

$$n = \frac{1}{C1\ article\ count + C2\ article\ count}$$

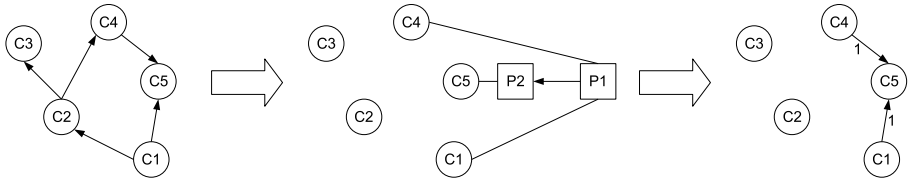

Fig. 1. New, weighed category links generation with our method. C denotes categories, P – Wikipedia articles.

The idea of the method has been depicted in Figure 1. The first part of the picture with nodes C1 to C5 describes the categories with unweighted links connecting them. The next step shows pages denoted by nodes P1 and P2 that belong to categories and are connected by an unweighted link. The last step describes processing these data and calculating weights for the links between categories based on connections between articles.

3 The data

Data used in this experiment are obtained in the form of database table delivered from the Wikimedia Foundation download page[1]. We used to generate new relations between categories following Wikipedia tables:

- Pages – containing page data, including page title and id
- Categories – containing category data, including category title and id
- Pagelinks – containing all the links between pages
- Categorylinks – containing original page category membership and category – category relations

For efficiency reasons in our experiments we analyze the Polish Wikipedia[2], which contains approximately five times less articles than the original English version. Estimated size of the data used in experiments can be portrayed in terms of row count for each table:

- Pages – 750 000
- Categories – 57 000
- Pagelinks – 23 400 000
- Categorylinks – 1 600 000

The first problem was processing the graph given in the form relational table, where relations have been stored in the form of article identifier (integer) – article title (string). The process of indexing these tables was unacceptably high so to reduce execution time a dedicated implementation was created which completely

[1] http://download.wikimedia.org/
[2] http://pl.wikipedia.org/wiki/

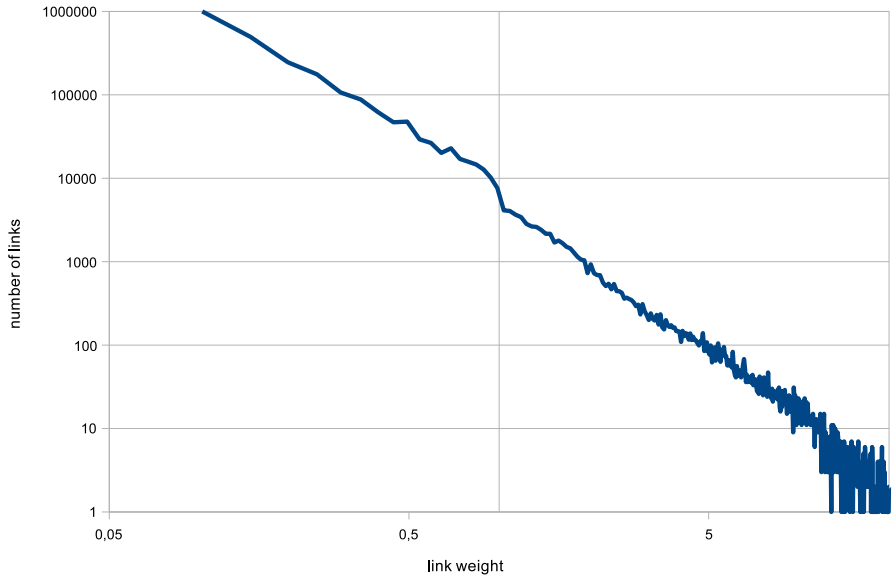

Fig. 2. Generated Category Graph edge weight distribution

disconnects data from database engine. The solution was a Java program, that reads the SQL dump files, in parts when necessary, into the memory, which requires Java Stack over 2GB. This method significantly reduced the impact of disk I/O operation times on the overall execution time. Furthermore, it allowed us to utilize the information about the data to use hashtables for rapid data access.

4 Results and evaluation

Using proposed approach generated 16 281 366 edges for the Category Graph from the Polish Wikipedia. An average edge weight value is 0.058, although only 20% (3 300 477) edges have a weight of 0.05 or more.

Figure 2 shows detailed weight distribution in logarithmic scale. As it was expected most of the edges have a very low weight so they do not bring any interesting information. During usage of the results for the enrichment of original the category system (section 5) they should be discarded.

In the following sections we will call the category graph containing new, generated edges the Generated Category Graph.

4.1 Original Category Graph coverage

The original Category Graph contains 79 582 edges which represent single level category membership. The full Generated Category Graph covers 77 113 of these

edges, that is 96.6%. If we reduce the Generated Category Graph by removing any edges with weights below 0.05, the resulting graph holds 3 300 477 edges, of which 70.9% (56 434) correspond with the original Category Graph edges.

To put these results in a perspective we calculate the probability Pr of generating an edge of the original Category Graph at random.

$$Pr = \frac{original\ category\ links}{possible\ category\ links}$$

There are 57 884 categories so there are 3 350 557 456 possible category links, therefore

$$Pr = 0.0024\%$$

Knowing that we can calculate the coverage of the original Category Graph by randomly generated edges. It would be 0.4% with a set of 16 281 366 edges and 0.09% with a set of 3 300 477 edges.

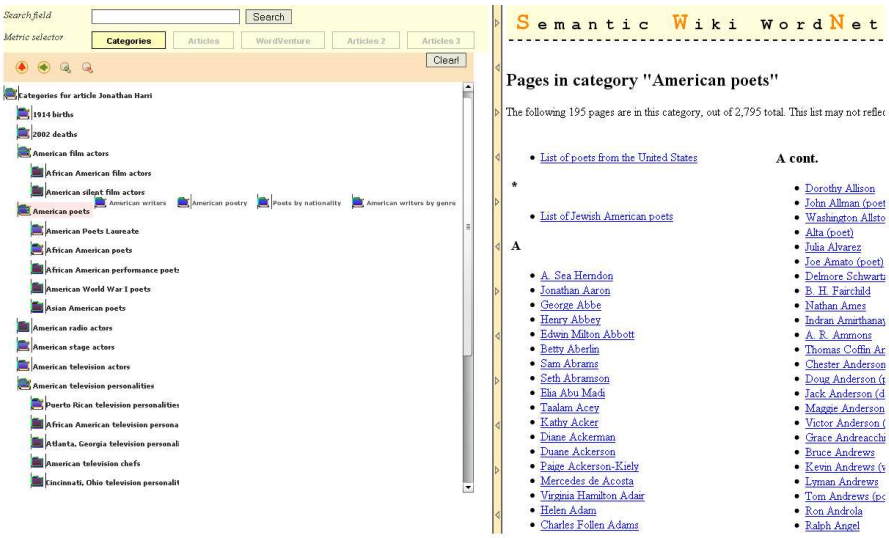

Fig. 3. An example of categories visualization

5 Visualization

Results of the presented here experiment has been used in or project aiming to find methodology for organization of textual knowledge. In project *Semantic Wiki WordNet*[3] we research algorithms that aims to improve knowledge organization in Wikipedia.

[3] http://swn.eti.pg.gda.pl

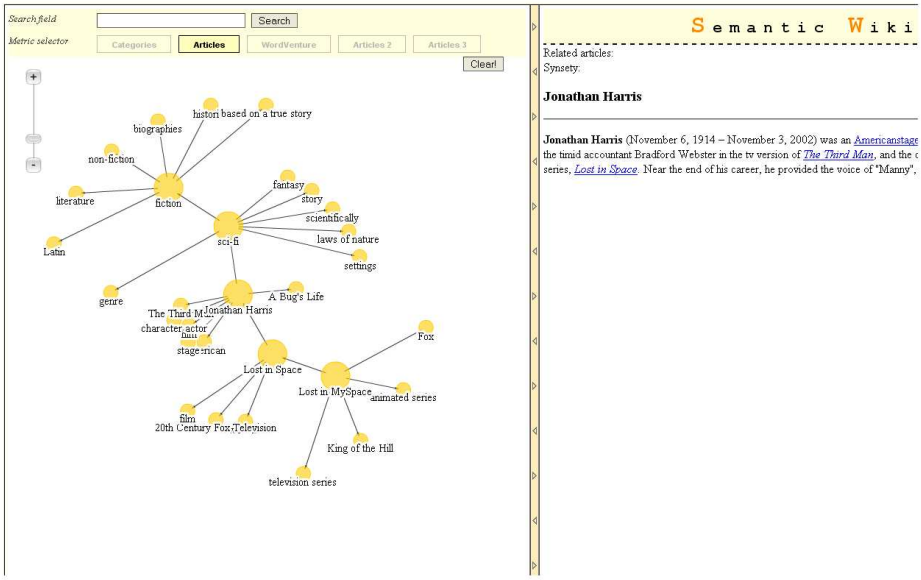

Fig. 4. An example of interactive Wikipedia visualization based on references between articles

We have used Generated Category Graph to recommend categories related to the article the user is interested in. Looking through this categories user may find other, valuable information. Categories related to the selected article and relations between them are presented using graphical web interface that allow to navigate thought Wikipedia categories in user-friendly way. The interface has been implemented using Flash component that allow to navigate over the Wikipedia Category System in a similar way as it is in the file system. The screenshot of the visualization where categories are presented as folders with multiple ancestors has been presented in Figure 3.

The system allows also to navigate between articles using the idea of the interactive graph. This functionality has been implemented using our component called Gossamer[4]. Gossamer is a general purpose solution that using Flash technology allow to visualize on-line large scale graphs. The component enables functionality to traverse the nodes of the graph using interaction with the user. The sample visualization has been presented in Figure 3 where hyperlinks between articles allow to traverse Wikipedia using graphical interface. Presented idea of visualization has also been used in project to develop WordNet in cooperative way [12].

[4] http://gossamer.eti.pg.gda.pl

6 Summary and future work

We were successful in developing a method that allowed us to identify additional relations between Wikipedia categories. In the article we present first result of our approach based on mining Article Graph, that seems very promising. After analyzing this first results we find some ideas of improvement of the method as well a new perspectives for research on mining machine readable knowledge from Wikipedia arise.

We plan to enrich representation of Wikipedia articles, that is now performed on links. We consider to introduce a representation of articles based on words and compute similarity based on their co–occurrences [13]. That should allow to process semantic relations between articles. It is also possible to introduce more sofisticated similarity measures based on article semantics exploiting additional, external information about language eg. WordNet[14] which is integrated with Wikipedia [15].

Acknowledgements

This work was supported by Polish Ministry of Science and Higher Education under research project N519 432338.

References

1. Viegas, F., Wattenberg, M., Kriss, J., Van Ham, F.: Talk before you type: Coordination in Wikipedia. In: Hawaii International Conference on System Sciences. Volume 40., IEEE (2007) 1298
2. Voss, J.: Measuring wikipedia. In: Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics. (2005)
3. Buscaldi, D., Rosso, P.: Mining knowledge from wikipedia for the question answering task. In: Proceedings of the International Conference on Language Resources and Evaluation. (2006)
4. Tan, A.: Text mining: The state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, Citeseer (1999) 65–70
5. Zesch, T., Gurevych, I.: Analysis of the wikipedia category graph for NLP applications. In: Proc of NAACL-HLT 2007 Workshop: TextGraphs. Volume 2. (2007)
6. Ponzetto, S., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: Proceedings of the national conference on artificial intelligence. Volume 22., Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2007) 1440
7. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD workshop on text mining. Volume 400., Citeseer (2000) 525–526
8. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34** (2002) 1–47
9. Majewski, P., Szymaski, J.: Text categorisation with semantic common sense knowledge: first results. Springer Lecture Notes in Computer Science, Proceedings of 14th Int. Conference on Neural Information Processing (ICONIP07) **4985** (2008) 285–294

10. Milne, D.: Computing semantic relatedness using wikipedia link structure. In: Proceedings of the New Zealand Computer Science Research Student conference (NZCSRSC 07), Hamilton, New Zealand. (2007)
11. Hao, P., Chiang, J., Tu, Y.: Hierarchically SVM classification based on support vector clustering method and its application to document categorization. Expert Systems with Applications **33** (2007) 627–635
12. Szymaski, J.: Developing WordNet in Wikipedia-like style. Proceedings of the 5th International Conference of the Global WordNet Association (2010) 342–347
13. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. (2007) 6–12
14. Miller, G.A., Beckitch, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University Press (1993)
15. Szymaski, J., Kilanowski, D.: Wikipedia and WordNet integration based on words co-occurrences. Proceedings of 30th International Conference Information Systems, Architecture and Technology **1** (2009) 93–103