See Also: Auto Generated Recommendations

Mislav Cimperšak, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mcimper1@ffzg.hr

Marija Tkalec, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mtkalec@ffzg.hr

Siniša Jovčić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sinisasanseverija@yahoo.com

Summary

Wikipedia is one of the most used encyclopedias today. It earned its status for more than one reason: easy accessibility, regular and quick updates and useful hyperlinks within the articles. Apart from linking to the other articles within the main body of an article, Wikipedia also uses a section called "See also" in which similar or related articles are listed.
The main idea behind this research is the creation of automatic recommendation for the "See also" section based on soft clustering of document similarities with the addition of the hyperlinks within the main body of an article as the observed objects. The research is conducted on the Croatian version of Wikipedia, with short articles omitted. In the article it was concluded that the procedure cannot be regarded as being successful enough for an unsupervised implementation on articles in Croatian Wikipedia.

Key words: Wikipedia, See also, document clustering, soft clustering

Introduction

Today Wikipedia[1] is a very reliable source of information which is accessible to everyone around the world. One of the biggest Wikipedia's advantages is that it

[1] Wikipedia, 2009.

is the most up-to-date online encyclopedia which means that its articles are in most cases most prompt of all articles from other online encyclopedias. People who want to extend their knowledge daily visit Wikipedia because of its quick and easy accessibility.
It is also necessary to mention Wikipedia disadvantages. The most talked about disadvantage is uncertainty of accuracy of information provided, but it should also be taken in consideration that articles with a lot of sources are in most cases more accurate. Because anyone can start a new article, topics that usually would be considered too obscure for a general encyclopedia are included leading to may very specific articles which are often very short.
Among many different sections, Wikipedia includes a section called "See also" which includes the list of similar or related articles to current article which urges users to continue browsing and reading articles on the page itself.
In classic encyclopedias entries often have further reading sections that list a handful of related articles or even entire books, making encyclopedias extensive references for scientific research. The same idea is built inside the Wikipedia using the connections between articles within the article text and the "See also" section on its own.
The main topic of our research described in this paper is the creation of an automatic recommendation system for the "See also" section based on soft clustering of documents.
The research was performed on 5,012 articles taken from the Croatian version of Wikipedia.

Thesis

Within the body of an article text Wikipedia's users can add connections to other articles on Wikipedia providing a useful way of finding out more about an unknown term within the article. The idea behind the research was that users on similar topics create connections to same articles. That would mean that by comparing two articles connections we could conclude how similar those two articles are. Since there are usually not that many connections per article based on which we could do the comparison, we compared the whole text with added extra weight on the article connections making them more important than the rest of the article text.

Database

10,100 articles were originally collected and run through a specific cleaning process. The cleaning process included the removal of all HTML tags (except connections to other Wikipedia articles within the text of an article), general parts of Wikipedia such as navigational components, articles with the same title, but in different languages. "See also"/"Related articles" parts were also removed since they would impose their Wikipedia connections to the overall re-

sults. After the cleaning process there remained 5,012 articles that we used as our main database.
Articles were collected between 12th and 27th August 2009 using the application "Wget" freely available online and preinstalled in most Unix-like operating systems. During the process of collecting some of the Wikipedia's webpages were omitted such as category pages since they contain only connections to articles without any real text on their own, further more editing pages, user pages, user's talk pages and Wikipedia help pages.
Articles shorter than 4,000 characters, after the above mentioned cleaning process, were discarded because they do not contain enough information needed for reasonable clustering.

Research

Based on a previous research[2] in which the optimal procedure for one-pass document soft clustering[3] was determined, we used tokens as vector features and document similarity threshold of 0.5.
We compared our results to human created connections to similar articles (where they were available) on Wikipedia itself taking in consideration the objectivity of selected connections (there are well known cases of private companies editing Wikipedia under the mask of ordinary users and adding connections to the article about their company and/or product creating for themselves a sort of free advertisement).
Connections within Wikipedia were treated as separate tokens which were given extra weight when comparing the articles.
Out of 5,012 articles, 509 clusters were created. Average size of a cluster was 14.12 articles per cluster and a single article was places in averagely 2.4 clusters.
Analyzing the final clustering results we note that clusters can be classified in three categories. First category includes clusters which have no real value, i.e. clusters which contain totally incoherent articles of quite different areas. Second category includes partly relevant clusters containing some articles of the same area and others of some different topic area. And finally the third category includes well-formed clusters whose articles are completely based around the same area and therefore it is obvious that they would be very good candidates for one's "See also" section.

Clusters with no real value

In a certain amount of cases, generated clusters were not usable, because the subject of those articles is in a completely different theme area. Explanation:

[2] Cimperšak, Tkalec, 2009.

[3] Jain, 1999.

because of the added weight on the connections to the other Wikipedia articles the algorithm grouped together articles that were at first glance not related, but were for an instance connected through a random number (whether a number on its own or a year within a date etc.) or through a country name as the case was in the example of cluster number three which contains four members.
There is an apparent case of total disconnection of articles placed in that cluster. The topic of article number 4929 is St. Peter, article number 4450 Saint-John Perse, article number 1697 General Staff of the Armed Forces of the Republic of Croatia, article number 1709 French Guiana and article number 3027 Marine mammals. At first sight, it's obvious that all of the above mentioned articles are not related and they are not covering similar themes. Another example of such a cluster was cluster number 11 which consists of three members where the topic of the first article are Eurasian Avars, the second is Psychology and the topic of the third article are birds.
With this term, clusters with no real value, we could also describe clusters which contain too many articles, such as clusters containing more than 30 articles, and there were a fair number of clusters of that type. Since our research was conducted on a "smaller" Wikipedia[4], we consider that it is unwise to take into consideration clusters of such enormous size, because it is impossible that articles inside those clusters are connected through the same topic, and hence most of the articles within the cluster are absolutely unconnected. As an example, we enclose cluster number 171 which contains 39 articles and cluster number 530 which consists of 201 articles from which is obvious that it is not possible that all of them are closely related through the same topic considering Croatian Wikipedia.
The number of useless clusters is unfortunately larger than anticipated which takes us to the conclusion that our algorithm for finding related articles isn't creating satisfactory results in this case.

Partially relevant clusters

Some articles within this kind of clusters are thematically related, while the remaining articles are not bound with the same subject or they don't involve the same or similar area. A good example of such a cluster is cluster number 529 where the subject of an article number 1908 is the Croatian Football Federation, the subject of articles 1909 and 1911 are Parliamentary elections, 1919 Orthography, 1914 Presidential elections, 1768 Croatian Academy of Sciences and Arts and finally 2816 Loyalist (American Revolution). Analyzing the example above, everyone can notice that one part of articles share a similar theme, because they talk about elections (1909, 1911, 1914), some about Croatia (1908

[4] By smaller we mean smaller in the amount of articles available on Croatian Wikipedia when compared to some others.

and 1768 directly and 1909, 1911, 1914 indirectly) while the remainder includes articles no more related to the same election theme.
Based on the given results, we can conclude that among all of the created clusters, partially relevant clusters are the most common.

Well-formed clusters
Clusters that explicitly contain articles connected to the same subject are too rare for our likings. Cluster number 432 which speaks about the Olympic Games is a very good example of a well-formed cluster. Articles placed in this cluster involve Olympic Games in Tokyo, London; Barcelona, Atlanta, Athena and Beijing. The subject of one article is Berlin and it contains connection to the 1936 Summer Olympic Games. To determine if the cluster is well formed, we evaluated our generated results to user created "See also" section of those articles in the Croatian Wikipedia. We observed that within the Olympic Games article there is a list of Winter Olympic Games created by users where all list items are shown as connections leading to articles where the subject of each article is the Winter Olympic Games hosted by another city.
Cluster number 357 includes articles whose common topic is football teams and every article belonging to that cluster describes a different team.
Cluster number 511 involves articles speaking of different varieties of Airbus airplanes. While the number of such articles is six, it also includes an article about a Boeing 747 for which is clear that the main topic is still an airplane. One of articles is about Ahmed Sékou Touré and it contains a connection to the article about France which was probably crucial when placing that article under the above mentioned cluster since the centre of the European air company Airbus is in Toulouse in France and articles about Airbuses also include connections to the article about France. When we compare our results to user created "See also" section in the article about Airbus 380 there are connections to articles about Airbus A350 and Boeing 747 (and to three more airplanes, but their articles were not in our database since they were shorter than 4000 characters) which demonstrates our initial concept.

Observations
During the course of our research we noticed that Wikipedia users (by users we mean contributing users) more often create connections on more general and, at first glance, more obvious terms such as dates (whether complete dates or just names of the months or years), geographical terms (cities, countries), objects of general use etc. For example, article number 4329 about Roman-Persian wars contains a great number of connections leading to articles about Iranian empires, Mesopotamia, Arab Muslim armies, Euphrates, North Africa etc. which are all general terms. More seldom they create connections to highly specific or specialized terms or terms less known in general public (such as a name of an

unfamiliar town), because they are not aware that the article about that term was even published.

Conclusion

Based on the results presented here, the procedure cannot be regarded as being successful enough for an unsupervised implementation on articles in Croatian Wikipedia. Unfortunately it would be necessary to manually inspect each generated recommendation.

Most likely the algorithm would be more successful in a strictly supervised encyclopedia where connections to all possible terms would exist and not just on those which are more familiar to an average Wikipedia user.

References

Cimperšak, Mislav, Tkalec, Marija. Utvrđivanje optimalnog postupka za raspršeno grupiranje jednim prolaskom. Zagreb; Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu. 2009

Jain, Anil K., Murty M. Narsimha, Flynn, Patrick J. Data Clustering: A Review. ACM Computing Surveys. Vol. 31, 1999. No. 3; 264-323

Ljubešić, Nikola; Agić, Željko; Bakarić, Nikola. Document Representation Methods for News Event Detection in Croatian. Zagreb : Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu, 2008. 79-84

Wikipedia. Wikipedia. http://en.wikipedia.org/wiki/Wikipedia (19th August 2009)