Management of Textual Data at Conceptual Level

Julian Szymański

Department of Electronic, Telecommunication and Informatics
Gdańsk University of Technology, Gdańsk, Poland
Email: julian.szymanski@eti.pg.gda.pl

Abstract: The article presents the approach to the management of a large repository of documents at conceptual level. We describe our approach to representing Wikipedia articles using their categories. The representation has been used to construct groups of similar articles. Proposed approach has been implemented in prototype system that allows to organize articles that are search results for a given query. Constructed clusters allow to show directions in which user may continue the retrieval process to narrow search results.

Keywords: text representation, information retrieval, clustering

I. Introduction

Retrieving relevant information requires knowledge of the search phrase that index this information. The user usually doesn't know all aspects of the retrieved information which may be useful for detailed specification his requirements. Thus the search engines provide some hints that may improve the search quality. Eg. Google suggests additional phrases that may be added to the search phrase also the search engines provide functionality of retrieving relevant pages to a given one. Other approaches suggest directions where user may continue his or her search based on clusters of similar data.
In the article we present our approach to text representation and show how employing clustering methods select conceptual directions in which user can continue the search. We demonstrate our prototype system that is able to cluster Wikipedia articles and using its categories provide directions that allow to narrow search results.

II. Text representation

Typical approach to text representation is based on features extracted from the text. The features allow to construct geometrical space where documents are treated aspoints coded as features representation vector. The **V**ector **S**pace **M**odel (VSM) [9] allows to compute similarities between documents and is a basis for application data mining algorithms.
There are two main approaches to obtain features from the text:

- The first one is the usage words that appear in the text. The method does not take into account the order of the words that appear in the text and it is called **B**ag **o**f **W**ords.
- The second method utilize relations between documents that are provided explicite. In this approach bibliographic references can be used. If *html* documents are considered the referential space may be constructed using hyperlinks.

The document representation allows to put document into n-dimensional feature space and perform computations on the text. Providing good features is a basis for obtaining good results of automatic text processing. The main problem with the features based on words and references is that they do not capture semantics of the text. If similar documents are described with different words these methods cause that the documents will be computed as different ones. The next issue is that the features that use references produce sparse representation vectors. Additionally the features based on words produce vey high dimensional spaces. There were made many modifications of these approaches, eg.: introduction of correlations between words in Generalized Vector Space Model [12].

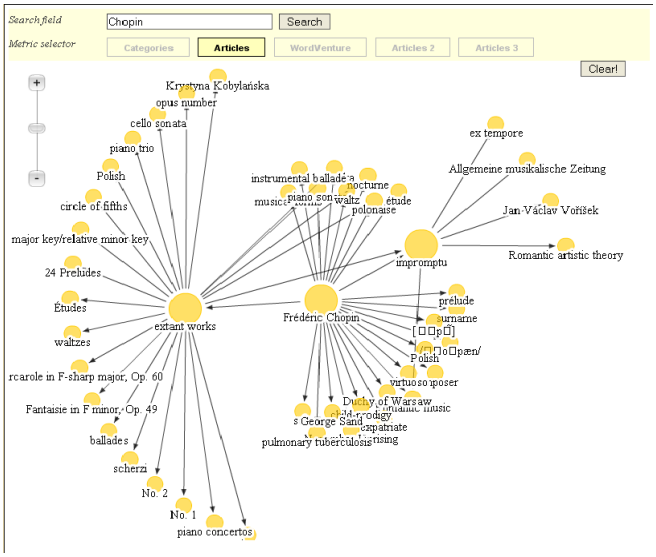

Figure 1 Link structure visualization used for Wikipedia articles

The other improvement has been made combining representation based on words and links in the form of Contextual Network Graphs [1].
The next problem with the representation based on words is that it requires to apply **N**atural **L**anguage **P**reprocessing (tokenization, stemming, part of speech tagging etc.), which may introduce additional noise into the data. It is mainly caused by imperfectness of the algorithms for NLP. Because of that in our approach we decide to use references between articles that are easily extracted and are known to be high quality features.
In Figure 1 we present visualization of Wikipedia link structure that has been used to construct representation space. The links form a wide graph where nodes stands form articles and the directed edges of the graph represent references between articles. The visualization of this graph has been made using our component called gossamer[1]. The component enables functionality for operating on large graphs displaying only part of them. Traversing across the nodes allows to walk through the whole graph.

II.II. Conceptual representation

For organizing its content Wikipedia use the system of the categories. This system allows to find articles related to the given one. We implement a graphical component that allows to navigate across hierarchy of the categories in file system - like style. The example of this approach is presented in Figure 2. The component allows to navigate across the categories displaying its hierarchical structure.
What should be stressed here is the fact that categories may have more than one parent node thus they do not form exact tree. It requires to implement additional features that allow to display multi parent nodes for a selected one.
Usage of Bag of Words and link representation may fail when no abstract but only very specific information is included in the text. This fact implies necessity for providing additional knowledge to capture semantic of the text [6]. Using categories that combine groups of documents we were able to introduce additional information to the articles computational representation. Employing categories we group the references between articles into the sets of similar ones. The idea of the representation is depicted in Figure 3 where articles that link to the articles from the same category are bound together. Thus the representation vector of the article may be enriched with categorical information.
As it is shown in Figure 2 the categories in Wikipedia are related one each other. Employing this fact additional information about similarities of the articles on higher levels of abstraction may be computed. We introduce additional categories to the article representation vector with the weight calculated according to formula $e^{-\|dist/2\|}$, where *dist* is distance between categories. *dist* is computed as a number of transitions from the category node to which article link is related explicitly to selected upper category.

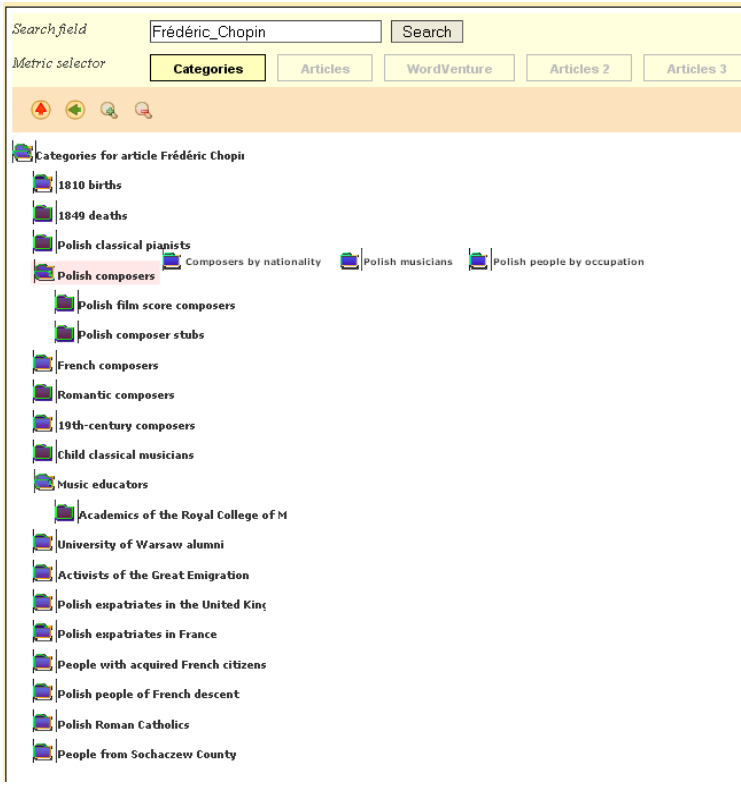

Figure 2 Hierarchical navigation across Wikipedia category structure

III. Density based clustering

The computational representation of articles may be used for computing the text. One of the possible applications is the usage of unsupervised methods of data organization to introduce groups of similarities. This process called clustering may be performed in several ways. One of the most effective approaches is clustering based on analyzing densities of the data points. The algorithm works by expanding clusters to their dense neighborhood thus it has several advantages:

- The main is it is able to produce clusters of different shapes,
- It is able to deal with noise in the data,
- It does not require to define a priori the number of clusters, instead it uses two parameters – ε that describes analyzed distance on the neighborhood points and *m* – the number of points,
- The application of density based approach to a text is a technique that is known to provide good results.

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

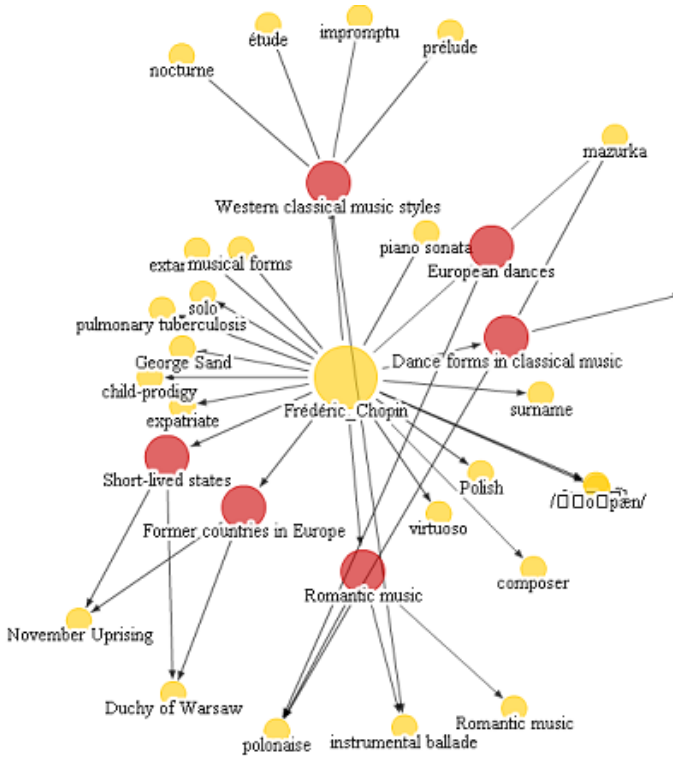

Figure 3 Graphical presentation of Conceptual representation for sample article Frederic Chopin

As we operate on high-dimensional data its processing cause several problems called curse of dimensionality [7]. The main issue here is the fact that while increasing dimensions distribution of the average distances concentrates around their expectations [4]. The next problem with processing natural language data is that the vectors of the representation are sparse. Due to that we use to calculating distances cosine similarity measure which is known to work well for high dimensional space vectors [10].
The well known algorithm for clustering based on data densities is DBSCAN [2]. The main concept of the algorithm is the notion of density reachability that is defined as an ability to reaching the point q from p moving across the data using circle of a given radius ε and having given number of points within it. The algorithm starts from the randomly selected point and gathers the points that are density reachable. Than it expands the points set joining the other points that are density reachable from the boarder points. The points that belong to ε surroundings but they do not contain enough neighbors are marked as noise.
This algorithm has been applied to the articles represented using the method described in section II.II.

IV. Evaluation measures

Evaluation of clustering may be performed using two main groups of criteria [8]:

Internal measures. Is the evaluation without external knowledge cohesion and distance of clusters are validated here. There are many indexes that stress on different aspects eg.: *Jaccard coefficient, Rand index* [3]. The internal metrics are the objective function of a clustering algorithm that can be calculated as:

$$\Phi(C) = \frac{1}{|C|}\sum_{c_i \in C} |c_i| \times sim(c_i, c_i)$$

where *sim()* is similarity function between clusters c_i. Internal metrics are reported to be the best measures for comparing clustering results on the same data collection. However one has to keep in mind that these measures analyze only formally defined structural aspect of the clusters but not their semantic content.

External measures. The second group of metrics allows us to evaluate the received results according to human made decisions. One of the informal measures is to collect feedback from users of a clustering system in the form of questionnaires. Another type of external measure is formal metric based on a relevance set. The most popular are Precision (P), Recall (R) combined into F-measure and Purity.
Precision is the percentage of retrieved documents that are relevant (that belong properly to the cluster):

$$P = \frac{Number\ of\ relevant\ documents}{Total\ amount\ of\ documents}$$

Recall is defined as the percentage of relevant documents that were grouped:

$$R = \frac{Number\ of\ relevant\ documents}{Total\ amount\ of\ relevant\ documents}$$

F-measure is a composition of Precision and Recall (weighted harmonic mean) and keeps a balance between them [5] :

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

where β (1,∞) is a weight coefficient. For β = 1 F-measure balances P and R. By increasing β we put emphasis on Precision. Most common values for β are 1, 3.
The next external measure is Purity. It is the normalized measure of correctly assigned articles to a cluster. It is computed by selecting the number of correctly assigned documents and dividing it by N (total amount of documents), when each cluster is assigned to the class which is the most frequent in the cluster:

$$Purity(\Omega, C) = \frac{1}{n}\sum_k max_j |\omega_k \cap c_j|,$$

where Ω represents clusters set, C is a set of classes, ω_k is k-th cluster and c_j is j-th class.

IV. Results and application

The approach for clustering based on DBSCAN (described in section III) Wikipedia articles using representation described in section II.II has been implemented in the form of web portal. The project is still under development and its actual version is available under http://swn.eti.pg.gda.pl/UniversalSearch/. The system allows to enter the search phrase and retrieve articles that contain this phrase. Than the system organizes these articles into groups according to defined similarity measure (here cosine of conceptual representation). As we used categories for articles representation the computed groups bind together articles that are conceptually similar and show possible directions where user may continue the search process.
We analyze for the 7 test phrases how the retrieved and clustered articles are conceptually similar. The evaluation has been performed revising the search results for the given query and manually judging whether articles in a given cluster are similar to each other or not. The results have been shown in Table 1.

Search phrase	Number of clusters	Φ()	F1	F3	P
kernel	10	0.81	0.82	0.85	0.87
mathematics	27	0.62	0.68	0.66	0.61
Europe	32	0.68	0.68	0.69	0.71
elk	4	0.94	0.95	0.94	0.93
claud	4	0.73	0.70	0.69	0.68
brain	12	0.67	0.67	0.71	0.70
wolverine	7	0.91	0.92	0.94	0.93

Table 1 Results of clustering Wikipedia articles for the test keywords.

The results seem promising and we decide to implement the method in the form of web portal. What can be seen is that quality of the results is better when more specific keywords are given. For general terms more articles are retrieved and clusters tend to form broader conceptual areas.
The snapshot of the application has been presented in Figure 4. The interface allows user to enter the search phrase, retrieve the Wikipedia articles that contain specified word and then organize them into groups. Due to efficiency reasons user may also select the number of retrieved articles that are taken into computations. By now only Polish Wikipedia is supported, the next step is to implement our approach for other languages.

V. Future directions

In the article we present approach for retrieving Wikipedia articles using representation based on categorical aggregation of the references.
The clustering algorithm we used depends on similarity measure that is used to compute distances between data points. Beside text representation it is very important element to obtain good clustering results. In future we plan to perform computations based on local distances. This may capture more precise relations than global cosine distance which we use now in our application. The main idea of development of local similarity measures is to sort features in the article description vector and compute the similarity of the articles according to local features similarity.
Introducing different similarity measures should provide different ways of organization thus stress different aspects of the interrelations of the articles [11].
We also plan to develop our online application and extend it to other than Polish.

Acknowledgements

The work has been supported by the Polish Ministry of Science and Higher Education under research grant N519 432 338

References

[1] M. Ceglowski, A. Coburn, and J. Cuadrado. Semantic search of unstructured data using contextual network graphs. *National Institute for Technology and Liberal Education*, 10, 2003.
[2] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd international conference on knowledge discovery and data mining*, volume 1996, pages 226–231. Portland: AAAI Press, 1996.
[3] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: part I. *ACM Sigmod Record*, 31(2):40–45, 2002.
[4] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Verlag, 2007.
[5] C.D. Manning, P. Raghavan, H. Sch\"utze, and Ebooks Corporation. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, UK, 2008.
[6] P. Matykiewicz, W. Duch, P. Zender, K. Crutcher, and J. Pestian. Neurocognitive approach to clustering of PubMed query results. *Advances in Neuro-Information Processing*, pages 70–79, 2009.
[7] V. Pestov. On the geometry of similarity search: Dimensionality curse and concentration of measure. *Information Processing Letters*, 73(1-2):47–51, 2000.
[8] Magnus Rosell. Introduction to text clustering, 2008.
[9] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[10] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.
[11] J. Szymański and W. Duch. Dynamic Semantic Visual Information Management. *Proceedings of the 9th International Conference on Information and Management Sciences*, pages 107–117, 2010.
[12] S.K.M. Wong, W. Ziarko, and P.C.N. Wong. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25. ACM, 1985.

Julian Szymański received the MSc and PhD degrees from Gdańsk University of Technology in computer science and MSc from the Nicolaus Copernicus University in philosophy; he works as assistant professor at Gdańsk University of Technology. The main areas of the research are cognitive language modeling with application to information retrieval.

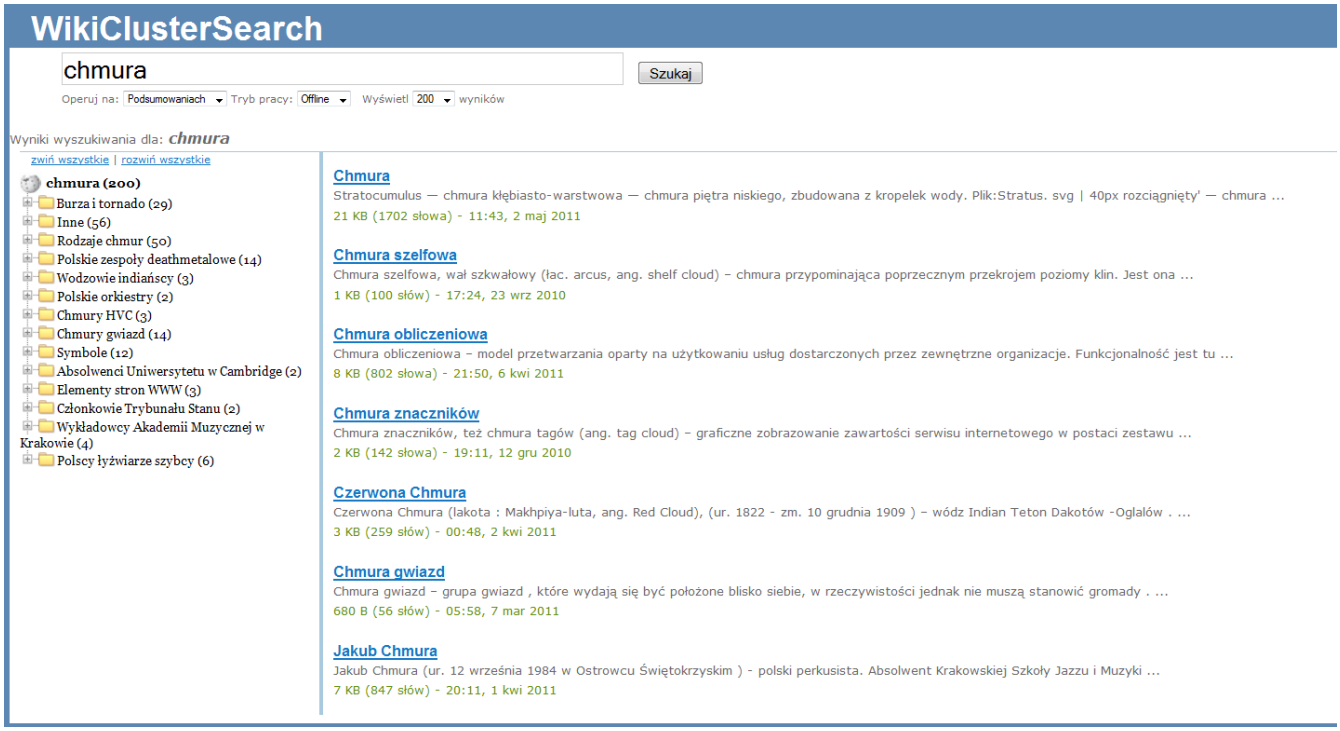

Figure 4 Snapshot of the application for clustering Wikipedia search results