frontiers in Robotics and AI

METHODS
published: 31 January 2017
doi: 10.3389/frobt.2016.00078

Spectral Clustering Wikipedia Keyword-Based Search Results

Julian Szymański and Tomasz Dziubich*

Department of Computer Systems Architecture, Gdańsk University of Technology, Gdańsk, Poland

The paper summarizes our research in the area of unsupervised categorization of Wikipedia articles. As a practical result of our research, we present an application of spectral clustering algorithm used for grouping Wikipedia search results. The main contribution of the paper is a representation method for Wikipedia articles that has been based on combination of words and links and used for categoriation of search results in this repository. We evaluate the proposed approach with Primary Component projections and show, on the test data, how usage of cosine transformation to create combined representations influence data variability. On sample test datasets, we also show how combined representation improves the data separation that increases overall results of data categorization. To implement the system, we review the main spectral clustering methods and we test their usability for text categorization. We give a brief description of the system architecture that groups online Wikipedia articles retrieved with user-specified keywords. Using the system, we show how clustering increases information retrieval effectiveness for Wikipedia data repository.

Keywords: documents categorization, text representation, spectral clustering, information retrieval, Wikipedia, human–computer interaction

OPEN ACCESS

Edited by:
Zdzislaw Kowalczuk, Gdańsk University of Technology, Poland

Reviewed by:
Dominique Chu, University of Kent, UK
Xinlei Chen, Carnegie Mellon University, USA

****Correspondence:***
Julian Szymański
julian.szymanski@eti.pg.gda.pl

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *11 September 2015*
Accepted: *19 December 2016*
Published: *31 January 2017*

Citation:
Szymański J and Dziubich T (2017) Spectral Clustering Wikipedia Keyword-Based Search Results. Front. Robot. AI 3:78. doi: 10.3389/frobt.2016.00078

1. INTRODUCTION

The categorization of documents is important task in the domain of automatic knowledge organization. This approach is based on the computation of similarities between objects, which finds many applications in the areas where a text should be analyzed. For example, in the Information Retrieval domain, Manning et al. (2008) introducing particular similarities between text (documents) allows to build structures that organize the documents repository and thus improves searching for relevant information. In our research, we propose a method for improvement of document retrieval within the Wikipedia repository. The approach is based on spectral clustering, and it employs mixed representation based on content and references.

The representation method of the data is crucial for achieving good results of it processing (Szymański and Duch, 2010). In this research, we propose a method for text representation based on a combined approach of words and references (links) between Wikipedia articles. We show on visualizations made with Principal Component projections, how the introduction of combined representation improves separation of the data that help to differentiate articles which belong to different categories. Cumulative analysis of components variability shows that combined representation allows us to represent the data with a much smaller number of features than using single representations, which considerably improves the efficiency of data processing.

To build a system presented in this paper, we evaluate the spectral clustering methods applied to identification of groups of thematically similar Wikipedia articles. The selected clustering method allows us to organize results of keyword-based searches into a hierarchy that increases the effectiveness of information retrieval in this repository. The evaluation of spectral clustering algorithms (Szymański, 2011a,b) has been based on datasets constructed from Wikipedia articles. The categories of these articles, made by Wikipedia's editors, have been used as relevance sets used for algorithms evaluation. The evaluation indicated the most suitable clustering method for our purposes. It has been implemented in the online application that organizes Wikipedia search results into clusters that represent groups of conceptually similar topics. Binding search results into clusters is an alternative approach for presenting large collections of data, where groups of similarities are used instead of displaying ranked list of articles. Thus, the approach allows the user to more effectively review the content of the search results.

The article is constructed as follows: after the introduction where we the describe motivation for our research, we describe our method for text representation, which has been applied to Wikipedia articles. On the test datasets, we show how combined representation improves the effectiveness of information processing. The combined representation has been evaluated using PCA projections. The selected clustering algorithm has been used in our system described in the last section. The paper is finished with a discussion of the achieved results, and proposes potential directions of future development.

2. SPECTRAL CLUSTERING

There is wide range of clustering algorithms that can be used for text categorization (Aggarwal and Zhai, 2012; Wu et al., 2013). One of the most efficient methods are spectral algorithms (Mall et al., 2013), and in our research, we perform evaluation of these type of approaches for text clustering. The main advantage of this algorithms is they do not suffer from the problem of local optima and can produce clusters of different shapes (Yang et al., 2011) not only convex as it is in case typical clustering algorithms.

The assumption of the spectral approach is that the analyzed data are represented by a graph, where nodes denote particular objects, and the edge weights describe their similarities. In that approach, clustering is based on cutting the graph using the methods of spectral analysis (Cvetkovic et al., 1995). In the recent years, this theory has been intensively developed (Von Luxburg, 2007; Han et al., 2011), especially in the direction of clustering algorithms, based on graph partitioning, where the most well-known are Shi–Malik (2000), Kannan–Vempala–Vetta (2000), Jordan–Ng–Weiss (2002), and Meila–Shi (2000).

It is known that the spectral clustering methods give high quality partitions (Ng et al., 2002) and have good, polynomial computational complexity (Kannan and Vetta, 2004; Vazirani, 2013). There are many variants of spectral algorithms (Jia et al., 2014). Mainly, they differ in the way of construction transformation space, where eigenvectors are calculated, and their further usage. What is common—they treat source objects as graph nodes that are transformed into d-dimensional space. This space is constructed using spectral analysis, and their essential clustering is performed. We can select the three main steps of spectral algorithms (Verma and Meila, 2003):

(1) Data preprocessing: At this beginning, these data are preprocessed into its computational representation. For clustering the text documents, the most popular method is the usage of a Vector Space Model (VSM) (Salton et al., 1975). In this approach, documents using a particular representation method (see section 3) are mapped into feature space, where further computations are performed. During the preprocessing phase, the text is smoothed: so-called stop words, special signs (such as punctuation marks) are removed from the content of the documents. Also, in this stage, the words can be turned into their basic form. The basic form of the word is extracted with usage tools such as stemmers and lemmatizers. The first tool is based mainly on a fixed set of derivation rules, the second employs additionally the data from the attached dictionary. Usually lematizers obtain better results (e.g., the word "*were*" will be converted to "*be*," while stemmer returns "*were*").
(2) Spectral mapping: This stage distinguished the spectral approach. Using the data from the first step, some normalizations and adjustments are performed (e.g., creation of Laplacian matrix), then the appropriate number of eigenvectors is calculated.
(3) Clustering: After the spectral projection, objects are divided into sets having the highest similarity. Depending on the method—we obtain hierarchy of the objects or we perform partition into flat subsets. This step can be completed using well-known clustering algorithms, e.g., k-means (Wagstaff et al., 2001) or is based on identification of special properties found in the data such as so-called spectral gap.

In our research, we describe and evaluate three approaches for text clustering using spectral methods (Szymański, 2011a,b).

In our experiments (Szymański, 2011a,b), we evaluate three spectral clustering algorithms: **Shi–Malik algorithm** (Shi and Malik, 2000) (SM), **Kann–Vempala–Vett algorithm** (KVV) (Verma and Meila, 2003), and **Jordan–Ng–Weiss algorithm** (Ng et al., 2002) (JNW).

3. TEXT REPRESENTATION

Humans understand a text, while machines at this time are unable to do that. Thus, to compute a text, a set of characteristic features is provided to a computer. The characteristics form a representation of a document and allow one to differentiate it from the others automatically. In information retrieval, two main approaches to documents representation are used (Bradshaw and Hammond, 2002):

(1) based on a text content,
(2) based on a document's relations with others.

The first one typically uses the analysis of word frequencies. The representation that employs words as features is called **B**ag **o**f **W**ords (BoW) because it does not take into account the semantics of the utterances, but is based only on frequencies of word occurrences in the documents. A severe limitation of this representation is that of not taking into account neither the words' order nor the simple grammatical constructions, but its application for simple classification and clustering tasks is known to work well. Other approaches to representation, based on text content, use analysis of letter distributions, *n*-grams or successive *n*-words (Damashek, 1995). This approach deals with some issues related to the usage of single words, but they do not change processing quality very significantly, and they are known to introduce additional problems, e.g., increasing dimensionality that usually leads to higher computational costs (Szymański, 2011a,b).

The second representation method employs the fact that the documents are often related to each other. Using references between documents, the representation of the particular document is constructed with its associations to the others. In this method, the document is described by other documents it is related to. Below, we describe these two methods more in detail. Then, we describe our approach, which is a combination of methods based on words and references.

3.1. Text Content

Typical approach to text representation based on its content employs words and using BoW they form representation space. This approach, called Vector Space Model (VSM) (Wong et al., 1985), allows one to compute document proximity using different measures, such as cosine or the reverse of Euclidean distances (Steinbach et al., 2000).

In that approach, particular feature obtains a weight that describes its descriptiveness for a given document. Typically weight is estimated for *n*-th term and *k*-th document *w* as a product of two factors: term frequency *tf* and inverse document frequency *idf* combined into one value $w_{k,n} = tf_{k,n} \cdot idf_n$ (Salton and Buckley, 1988). The term frequency is computed as the number of its occurrences in the document and is divided by the total number of terms in the document. The frequency of a term in a text determines its importance for document description—if a term appears in the document frequently, it is considered to be more important. The inverse document frequency increases the weight of terms that occur in a small number of documents. The idf_n factor describes the importance of the term for distinguishing documents from each other and is defined as $idf_n = \log(k/k_{term(n)})$, where k is the total number of documents and $k_{term(n)}$ denotes the number of documents that contain term n (Szymański and Duch, 2010).

3.2. Links

The approaches for automatic categorization based on references have been used in many domains. Especially good results have been shown for topic identification, e.g., in Mahdi and Joorabchi (2010).

Typically features extracted for text representation lead to high-dimensional spaces. Without preprocessing and dimensionality reduction (Korenius et al., 2007), the size of feature space is equal to size of words dictionary that is number of the distinct words found in documents repository. While large datasets are considered, operations on term-based representations are very computationally intensive.

A more compact way to create representation of text is the usage of references that appear between objects that are to be computed. For the scientific articles or books the bibliographical notes provides very useful information. If we process web pages as features hyperlinks can be used.

Feature space based on hyperlinks or citations may be constructed in several ways. In the simplest approach, each reference states as a new dimension and the document representation creates a binary vector, where 1 denotes the presence of the link (reference) to another document and 0 indicate its lack. Documents on similar topics tend to link to similar sets of other documents and cite the same references. Possible extensions of this representation involve frequency of references, various forms of weighting, e.g., based on the position of a link in the document, and the use of directed links (± 1 for links from or to the document). These modifications have not been considered here, only binary representations of references between articles have been used in the experiments (Szymański and Duch, 2010).

3.3. Combined Representation

One of the approaches that combine the proposed above representations using a reference context has been shown in Aljaber et al. (2010). The authors use the text surrounding the references to extend the text representation for document clustering. The proposed approach has been also used for text classification in the medical domain (Ortuño et al., 2013), where it showed an improvement of processing results in comparison to the typical content-based representations (Aljaber et al., 2011).

To describe our method for using combined representation, we show how the variability of the data with successive raw representations based on words and links change while it is reduced. We also provide 2D visualizations to show how different representations change the distribution of data categories (Szymański, 2011a,b).

In the experiments, we use Wikipedia dumps[1] to create the dataset for evaluation methods for automatic document categorization (Szymański, 2011a,b). The advantage of using Wikipedia data is the fact it is easily accessible and free from the restrictive licenses. As it offers variety of articles from different domains, it covers wide range of human knowledge. Also, very valuable feature in using Wikipedia is it offers categories made by humans which can be used for the evaluation of the results obtained after machine processing.

To test different methods of the representation, we select 419 articles divided into 4 categories. **Table 1** provides details of our test dataset: categories and the carnality of the articles they include, the size of the feature spaces respectively to used representation method.

[1]http://download.wikimedia.org.

To reduce feature spaces, we made additional processing for reducing the dimensionality:

- for the representation based on links, we remove the features related only to the one article (articles that are referenced only by one other article).
- for the representation based on words, we remove the words that appear only once in a one article.

This adjustment reduces considerably (approximately three times) the size of representation spaces. In **Table 1**, we denote it with ↓.

To present the data distribution, we show in **Figures 1** and **2** the 2D visualization of the datasets constructed for representations based on words and links, respectively. For the presentation we used projections based on two highest principal components obtained with Primary Component Analysis (described in section 3.3.1).

In the figures, left side plots shows the raw data, the right side presents the same data processed with cosine distance.

The cosine distance (calculated using the formula 1) is especially useful in information retrieval, where the data are sparse (Qian et al., 2004). It should be noticed the computation of the similarities and using as a representation the proximity matrix is an easy way for reducting the dimensions: if initially raw data have the sizes $n \times m$, where $n \gg m$ after calculating the similarities, we can use for analysis matrix of the size $n \times n$.

$$d(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \quad (1)$$

Comparing the visualizations shown on left and right hand in **Figures 1** and **2**, mapping the data using cosine distance into metric space produces distribution of objects that can be easier separated. Also, it should be noticed that the visualizations have been created in 2D which implies a strong dimensionality reduction and some relations between objects cannot be presented in this space. For the further computations, we used higher number of dimensions where the data can be represented more precisely (Szymański, 2011a,b).

TABLE 1 | The data used to demonstrate and evaluate approach for combining the representations.

Category name	Symbol and color	Number of articles	Number of features in text representation	
Philosophers	Magenta □	110	Links	Terms
Ethics	Green +	131	2,523	18,411
Logic	Blue ·	170	↓	↓
Epistemology	Black *	8	945	5,819

3.3.1. Dimension Reduction

The representations of the textual data are highly dimensional. If we plan to process the large repositories, due to efficiency reasons it should be limited, also such a modification can lead to noise reduction. One of the most well-known approaches for data reduction is to replace original features with artificial ones. In the **P**rincipal **C**omponent **A**nalysis (Jolliffe, 2002) approach, the new features are created as linear combination of original ones. The approach is based on computing eigenvectors of the matrix formed from features correlations. In PCA, selected number of eigenvectors, having the highest variance, is used to form new representation space. The space is formed by multiplication of original data with eigenvectors sorted in descending order of corresponding eigenvalues. Then, the less meaningful dimensions are rejected.

To approximate our original data within given margin of distortion, a selected number of eigenvectors need to be specified. In our experiments, we took as much of eigenvectors that the data variance is kept around 90%. **Figures 4** and **3** present the

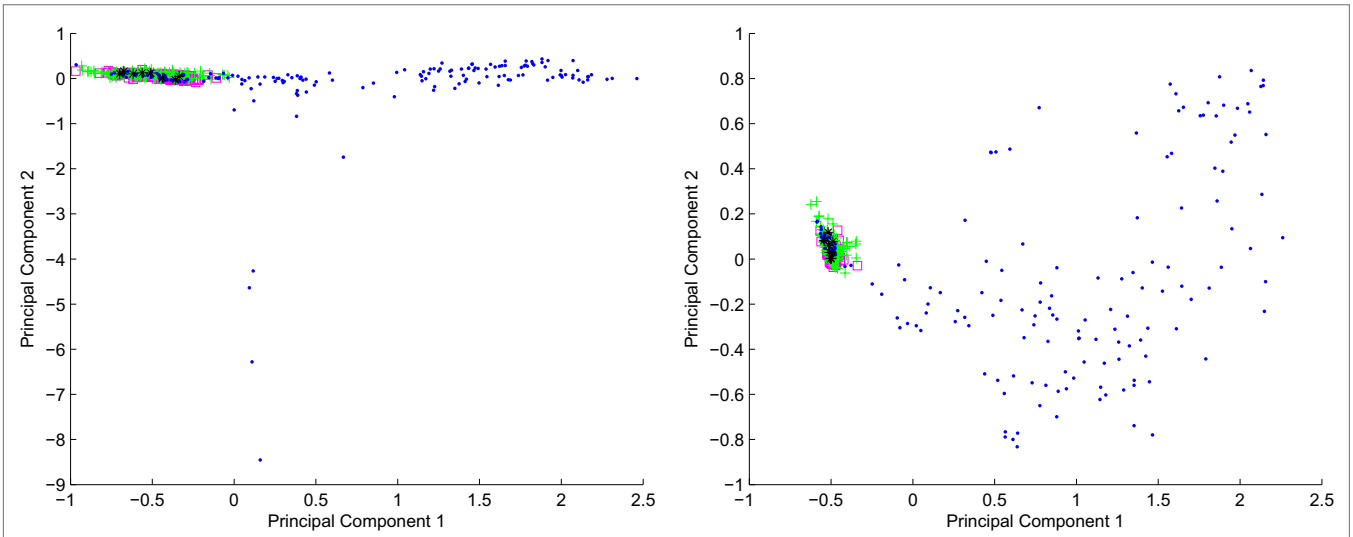

FIGURE 1 | Terms representation. 2D projections of the data: left hand—raw data visualization. Right hand—cosine distance-based data processing.

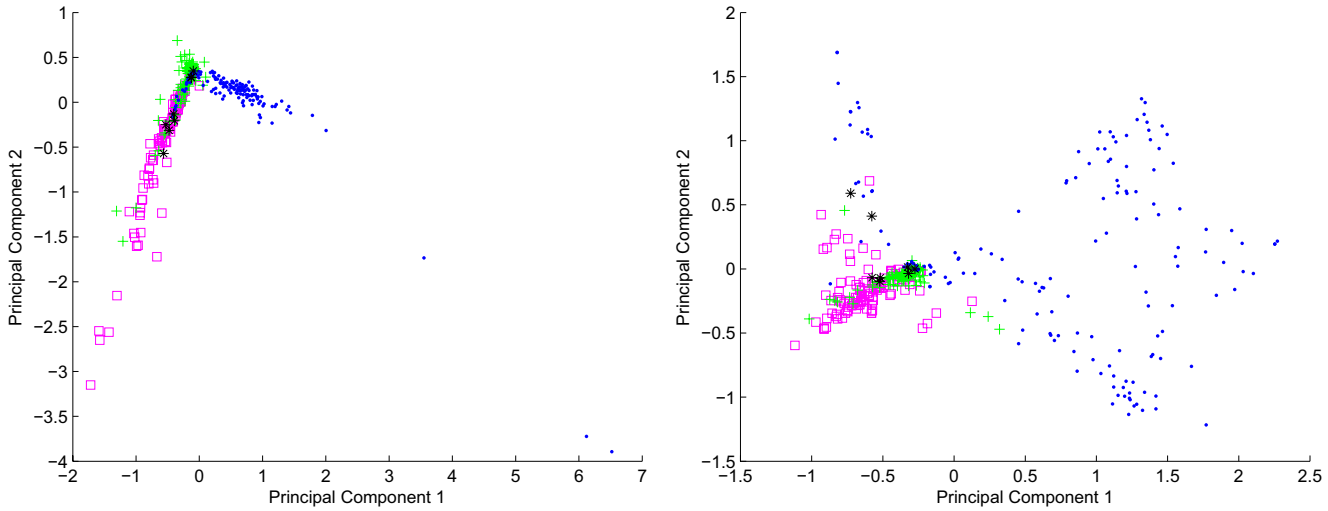

FIGURE 2 | **Representation based on links**. 2D projections of the data: left hand—raw data visualization. Right hand—cosine distance-based data processing.

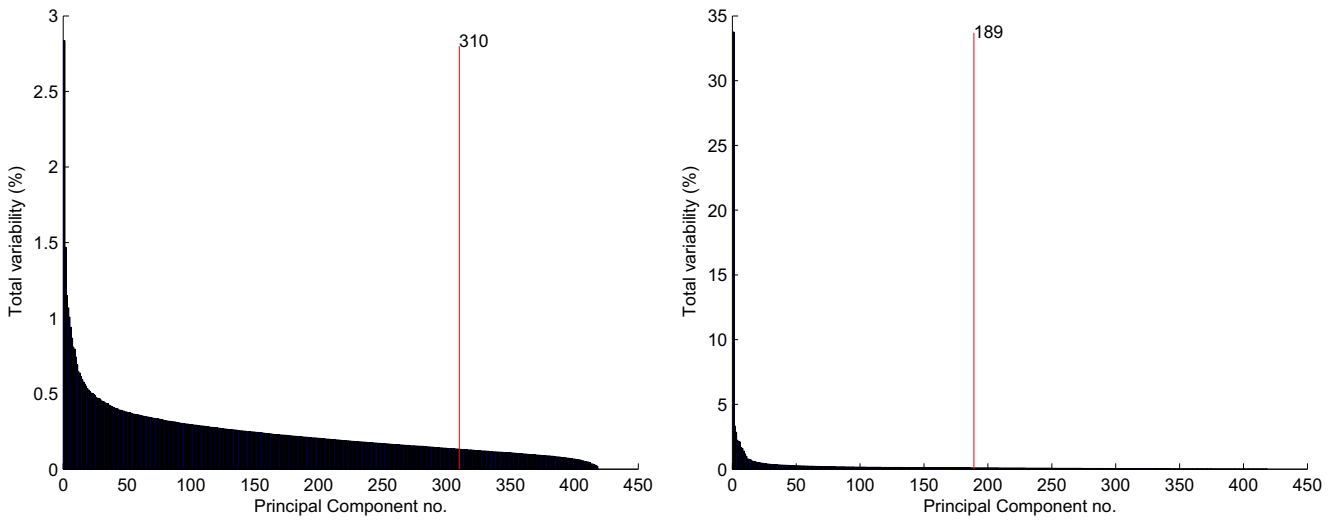

FIGURE 3 | **Representation based on terms**. % of variability: left hand—raw data visualization. Right hand—data processed with cosine distance.

% of variances for each of the components, respectively, to the used text representation. We provide also information about the number of components whose cumulative sum completes 90% of the variance.

In both cases of representations, usage of the cosine transformation for the data significantly reduce the number of the principal components required to complete 90% of the variance.

Because we are able to considerably reduce high-dimensional raw data, we can represent the articles using both types of representations jointly. To aggregate the two representations, we calculate the cosine distance within representation matrixes and perform reduction of the data based on identifying the most significant primary components.

For the data formed by combined representations, we show in **Figure 5** distribution of variances related to successive principal components. It should be noted the 90% of the variances have been completed using only 8 (!) principal components. It allows us to represent the raw joint data of the size 6,764 features (945 references and 5,819 terms) finally with only 8 dimensions, losing only 10% of data variability.

PCA projection obtained from the two highest components for the joint representation is shown in right hand picture in **Figure 5**. Comparing this figure with **Figures 1** and **2**, the transformed data looks to be easier separable than the original data.

What can be seen from the above example, the combined representation of words and links with cosine transformation,

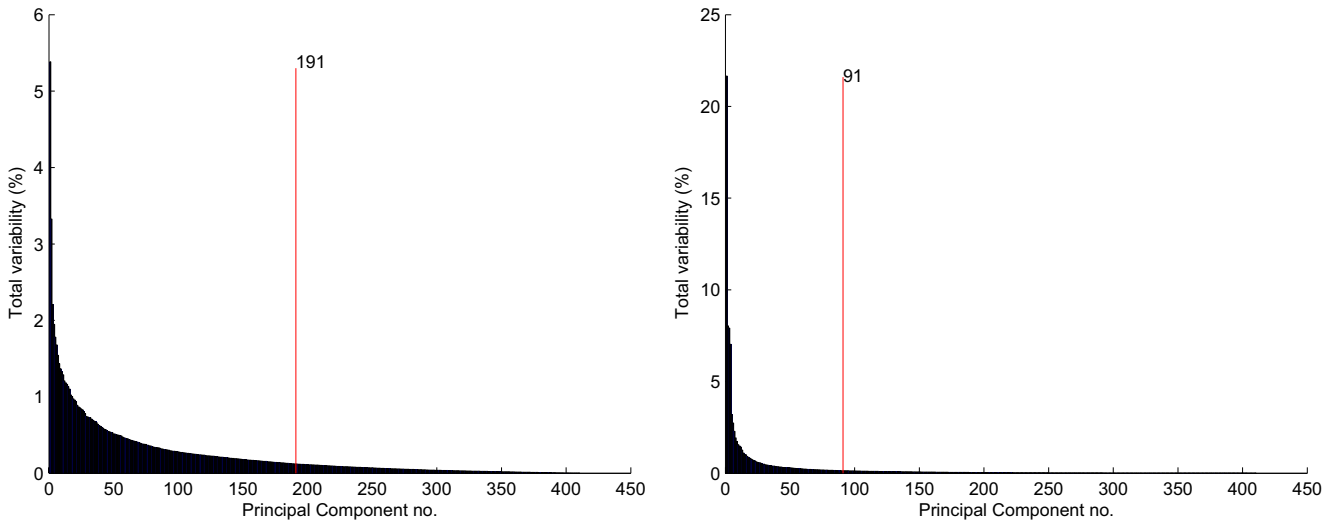

FIGURE 4 | **Representation based on references**. % of variability: left hand—raw data visualization. Right hand—data processed with cosine distance.

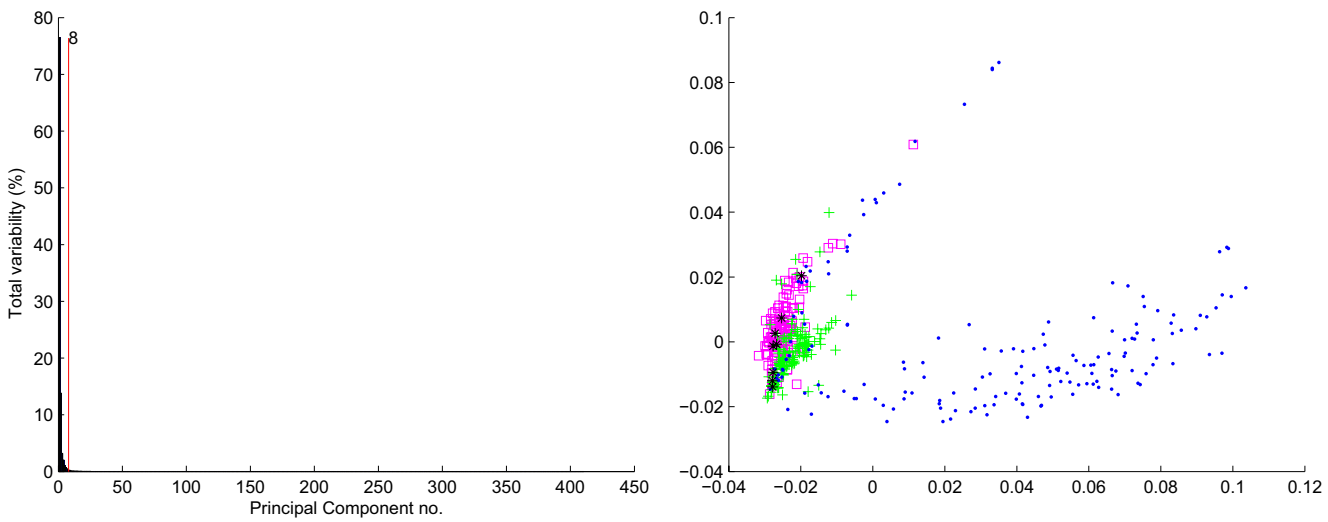

FIGURE 5 | **Joint representation**. % of variability: left hand—distribution of PCA variances. Right hand—data projection into 2D.

improves representation of documents. Also, the reduction of the dimensionality with the small number of principal components preserving high data variability, allows us to improve efficiency of computations (Szymański, 2011a,b).

4. EVALUATION OF COMBINED REPRESENTATION

Demonstrated in the previous section, usage of cosine transformation that combines two representations, has been evaluated measuring the influence of the particular representations on the categorization task. An objective of the experiment is that the better the results obtained using the same categorization method—the better the representation is.

Very popular approach to text categorization is the Naive Bayes classifier, e.g., Jiang et al. (2011). We use this widely studied approach to perform classification of the text within 10 test data packages. The evaluation dataset has been created in a similar way to the dataset presented in **Table 1**. The 10 packages for the experiment have been created using the data taken from articles from 10 arbitrarily selected Wikipedia categories. The categories for each data package have been selected randomly from the sets of categories having the same upper category. The process of building each of the data packages was based on

selection of the root category and then random selection of its 10 sub categories. As Wikipedia category structure is very irregular, it was decided, to select for each of the category in data package around 100 (±10) articles, all having similar length. This guarantees a uniform class distribution during categorization, and thus it does not bias the results.

In **Figure 6**, we present the results of performed classifications using different methods of representation based on words, links, and their combination. What can be observed almost in all data packages, combined representation leads to improvement of classification. Also, aggregated average results indicate that the combined representations work better than usage of single representation.

5. SELECTION OF CLUSTERING ALGORITHM

In our research, we aim at evaluation of the usage of unsupervised spectral methods for automatic text categorization. Thus, selection of the most suitable algorithm is an important issue. To evaluate clustering algorithms for text categorization, we constructed eight test data packages using a combined representation method based on words and links. The details of the experiments has been shown in Szymański (2011a,b).

The data for the experiments (shown in **Table 2**) have been generated using MATRIX'u software (Szymański, 2013). The software allows one to prepare Wikipedia content in a form that can be processed using machines. Among many functionalities, it allows you to select Wikipedia categories that narrow the set of articles and generate a set of representing features for them (according to a selected method of text representation). In the experiments, we use representations based on words and links but the application supports other approaches: based on references between the articles, suffix trees and common substrings (Grossi and Vitter, 2000), and information content computed by compression (Bennett et al., 2003). The application is available to download online,[2] and free for academic use. The comparative analysis of the text representation methods generated by the Matrix'u application can be found in Szymański (2014).

The observations we made from the experiments shown in Szymański (2011a,b) aiming at comparing spectral clustering algorithms indicate the most suitable for our task is JNW algorithm.

5.1. Application of the JNW Algorithm

Based on the JNW algorithm, we extend our system named WikiClusterSearch (WCS) (Szymański and Wegrzynowicz, 2011). As the algorithm is a partitioning approach to introduce the structure of clusters, we bind them using a typical Hierarchical Agglomerative approach (HAC) (Zepeda-Mendoza and Resendis-Antonio, 2013). In the approach to cluster binding, we use a modified linkage method: to average linkage (Krebs, 1999) method, where l denotes the minimal distance between two centroids, we add an additional factor ϵ. In the standard HAC approach, only the two nearest clusters are bind together. In our approach, we bound together the nearest clusters, and additionally join new elements if their distance is smaller than $\epsilon = \frac{l(n+1)}{2n}$, where n denotes the hierarchy level. This simple modification binds more actively clusters together on lower hierarchy levels. As we go higher into hierarchy structure, the method aggregates the clusters in a more similar way to HAC. Also, it allows us to form non-uniform structures, as sometimes it is required to put more than two clusters on the same hierarchy level (that is not possible using a standard HAC algorithm). The modification causes that the hierarchy is more bushy than

[2]http://kask.eti.pg.gda.pl/CompWiki/.

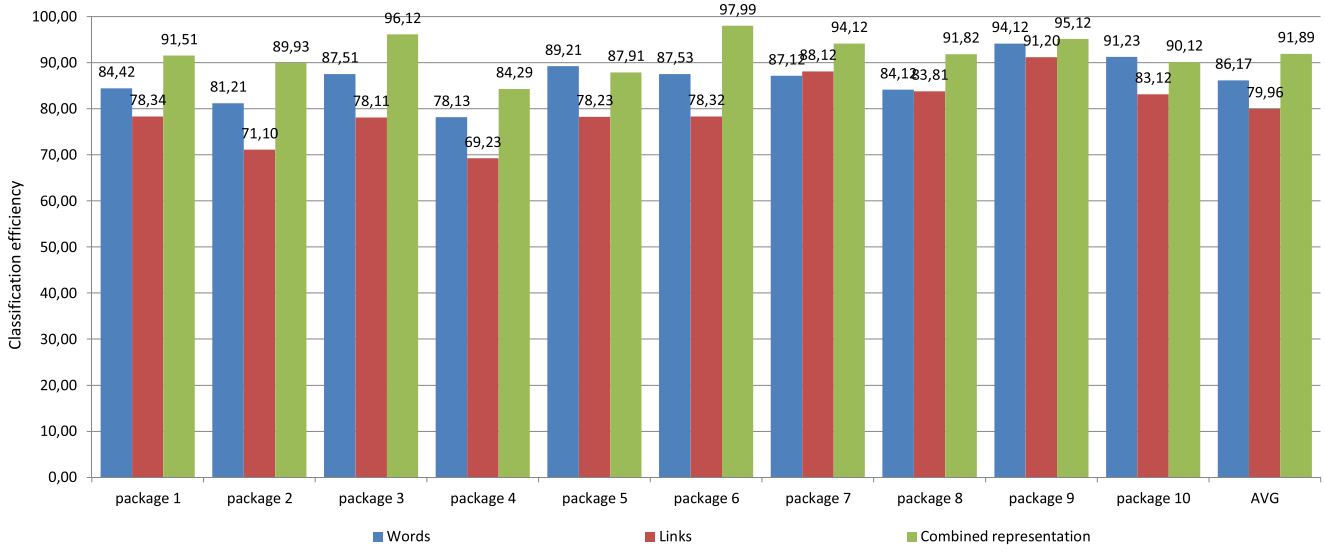

FIGURE 6 | Evaluation of combined representation.

TABLE 2 | Test packages.

No.	Name	n	l	N	k	d	P	Comment
1	Z01	575 (575)	67,099	2	18	1	3	2 distant categories: Distance_Education in Science_Experiments
2	Z02	1,157 (1,156)	323,901	5	35	1	3	5 distant categories: Caligraphy, General_Economics, Military_logistics, Evolution Analytic_number_theory
3	Z03	3,905 (3,903)	2,260,919	8	102	1	3	8 distant categories: Geometric_Topology, Epistemology, Rights, Aztec, Navigation, Clothing_companies, Protests, Biological_Evolution
4	Z04	3,827 (3,826)	3,195,963	2	204	6	4	Two distant categories at the same hierarchy level: Criticism_of_journalism and Corporate_crime
5	Z05	3,647 (3,644)	1,682,361	6	213	6	5	6 distant categories at high level of abstraction: DIY_Culture, Emergence, Military_transportation, Formal_languages, Geology_of_Australia, Computer-aided_design
6	Z06	4,750 (4,747)	2,568,378	9	289	6	5	9 distant categories at high level of abstraction: DIY_Culture, Emergence, Military_transportation, Formal_languages, Geology_of_Australia, Computer-aided_design, Special_functions, History_of_ceramics, Musical_thatre_companies
7	Z07	4,701 (4,701)	4,230,139	2	298	6	4	2 neighboring categories at high abstraction level: Computer_law and Prosecution
8	Z08	5,717 (5,716)	11,288,283	4	893	6	4	4 neighboring categories at high abstraction level: Impact_events, Droughts, Volcanic_events, Storm

n—number of nodes (in brackets—unisolated), l—non-zero elements in neighborhood matrix, N—number of upper categories, k—number of all categories, d—depth of category tree, P—number of categories to Wikipedia root category.

a regular, deep tree (produced by HAC), that is more natural for lexical data.

With the WCS system, the user may specify a searched phrase and retrieve articles containing it, then this set is organized into clusters on the fly. The approach based on clustering allows us to present thematic directions within a result set. One advantage of such an approach is better data presentation than using the lists of results. One easy extension based on clusters is a functionality for the user to continue his or her search by selecting a particular group to be extended, using additional similarity measures, and find information not indexed by specified keywords (Szymański and Wegrzynowicz, 2011).

For the efficiency reasons, for now only Polish Wikipedia is supported. The English version of Wikipedia one is approximately 5 times bigger, and it requires a little bit different approach to data processing. The efficiency requirements push us to introduce functionality where the user can specify the number of articles that will be processed. The results of the research presented in this paper has been used for implementation of the clustering module in our system for improving retrieval of Wikipedia information described in Deptuła et al. (2013).

The WikiClusterSearch we implement in C# in .NET 3.5 environment. It has a modular architecture, which has been presented in **Figure 7**. Successive elements are as follows:

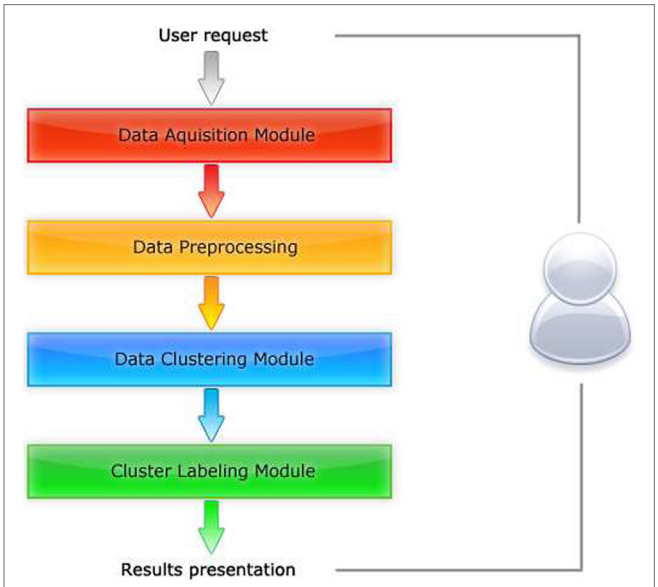

FIGURE 7 | Main data processing blocks in our system.

- Data acquisition module responsible for providing the data. In that module, we can specify whether we use off-line version of Wikipedia an we process the locally stored dumps. The other option is to process the data obtained form o-line repository thus we operate on the most actual data but for each query we need collect new data that can cause efficiency problems. To solve this issue, local cache has been implemented that need to be clean-up periodically to guaranty up to date information in the system.
- Data processing module creates a computational representation of the Wikipedia articles. In current implementation, we used the representation described in Section 3.3. Different other methods of text representation can be introduced here and the modifications in that module can allow to study the most suitable for obtaining the best results of automatic categorization.
- Data clustering module implements algorithms for grouping the data. In the implementation presented in this paper we employ the JNW algorithm. Further extensions plan to use clustering based on density analysis (Wang and Duo, 2007) and passing the messages (Frey and Dueck, 2007) within spectral spaces.
- Cluster labeling module to increase readability of the data adds labels to groups of the similar articles. Here, also different strategies can be used. In our implementation, we use the most frequent Wikipedia category that aggregates 90% of articles stored in a particular a cluster.

In **Figure 8**, we present a screenshot presenting an example of a system user interface. It shows clusters formed by the WCS system for articles retrieved from Polish Wikipedia for a sample query "*jdro*" (kernel). Created clusters form different conceptual directions in which a user may continue his or her search. The functionality of continuing the search can be extended by selecting a particular cluster and retrieving additional documents according to a provided similarity measure between documents in the cluster and the whole repository, which we plan to implement in the next system releases.

To evaluate the quality of clusters in terms of their commonsense, we use standard information retrieval metrics: precision and recall (Manning et al., 2008). Clustering precision is defined as the fraction of relevant text documents (Wikipedia articles) in the cluster, which is described in equation (2).

$$P = \frac{\text{Number of relevant documents in cluster}}{\text{Total number of documents in cluster}} \tag{2}$$

Recall is defined by the equation (3) as a percentage of all relevant documents that were aggregated into one cluster.

$$R = \frac{\text{Number of relevant documents in cluster}}{\text{Total number of relevant documents for cluster}} \tag{3}$$

F-measure is a composition using a weighted harmonic mean Precision and Recall values defined by equation (4).

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \tag{4}$$

where $\beta \in [0, \infty)$ is a weight coefficient. For $\beta = 1$, F-measure balances P and R. By increasing β, we put emphasis on precision. Most common values for β are 1, 3, and 5.

Evaluation of the system quality for a particular query has been performed for each of the cluster separately and final measure per query is a normalized averaged value achieved for all clusters. It should be noticed that calculation of "Total number of relevant documents for cluster" is a particularly time consuming task, as it requires one to look through all documents in the repository. Because of that, for each of test queries we extend the result set using query synonyms taken from WordNet. As articles in Wikipedia belong to the categories, we can select thematic domains of retrieved information. The articles from the categories indicated by the articles from the result set have been used to extend the original result set created with keywords. Additionally, we extend this set by adding articles related with them by hyperlinks. This extension significantly increases the number of articles that have to be analyzed for each query but it keeps their thematics relatively related to the, specified by the user, keywords. Thus, we do not need to look through the whole articles repository to create relevance set. During the evaluation process set of "Total number of relevant documents for cluster" we could create only according to retrieved documents but it

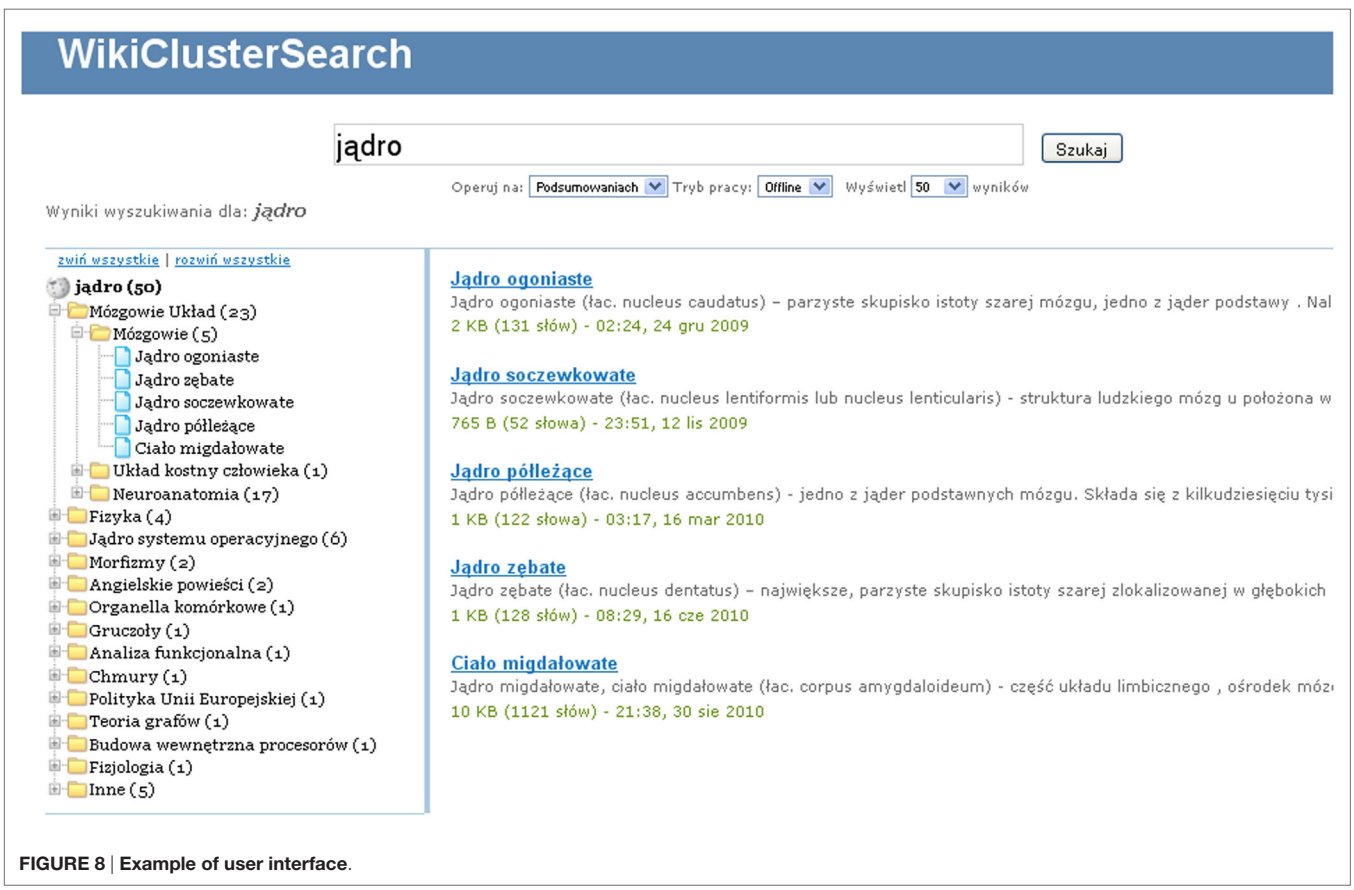

FIGURE 8 | Example of user interface.

would not give us information about misretrieved items. Also because of the fact, we perform retrieval using different number of items in result set used for clustering, this set should be review again for each search, that would make it hard to repeat the experiments. It should be noticed that the construction of relevance set ("Total number of relevant documents for cluster") in the proposed way causes obtaining lower values of recall measure. It is caused by the fact the relevance set is more detailed than the one created only by the search phrase.

In **Table 3**, we provide results of evaluation of clusters quality. They have been computed for two different parameters: variable number of articles that have been used for creating clusters and different values of parameter k indicating required number of clusters. The ϱ value indicates the ratio between articles and clusters. What can be observed, increasing the number of processed articles increases recall value, which seems to be natural as we obtain wider set to be processed. The advantage of the proposed clustering method is not decreasing the precision value, while increasing the number of processed elements. The high values of the precision measure indicate that the algorithm constructs a cluster that is thematically coherent. Also, we can observe that an increase in the number of clusters (k) leads to an increasing recall value. Presented results indicate the better retrieval is when ϱ ratio (articles to clusters) is smaller. It should be noticed, in our evaluation we do not test the quality of clusters hierarchy. In that case, if for a large number of articles we specify a high number of clusters the results will be characterized by good F-measure values, but there will be many cases where clusters should be bound together.

5.2. Further Directions

As we mentioned earlier, we plan to develop clustering algorithms for text documents and test approaches based on densities (Kriegel and Pfeifle, 2005) computed in spectral spaces. This approach should allow us to break through the problem of creating clusters only having convex shapes, as well as eliminating the requirement for k parameter specification.

We also plan to run our software for English Wikipedia that requires us to introduce some improvements into system architecture, as the scale of the data requires a more effective approach for storage and processing. The long-term goal is to join the method of retrieving the information based on clusters of Wikipedia categories with a classifier (Szymański, 2010) that allows us to categorize linear search results returned by a search engine into these categories. Currently, we are developing a new implementation that offers an extension of the clustering algorithm with functionality of narrowing search result by interaction with the user and extending the clusters' content using text similarity measures. The new version of the system is available at http://kask.eti.pg.gda.pl/BetterSearch, and it supports English Wikipedia.

We also plan to perform experiments with large scale clustering—it is on the whole Wikipedia. Here, instead of performing clustering within limited dataset (with, specified by the user, keywords), we plan to compute the whole Wikipedia. The size of the dataset in such experiments has to be run in a parallel environment, and it will be performed offline. The experiments, using a well tuned clustering algorithm would allow the improvement of the category system of Wikipedia finding missing and wrong assignments of articles to categories.

As text representation is crucial for obtaining good results, we plan to develop methods that are able to capture elementary semantics. We plan to introduce to the system some background linguistic knowledge. A mechanism inspired by cognitive theories of a language (Duch, 2009) such as the brain process called spreading activation (Collins and Loftus, 1975) should allow us to add concepts not explicitly present in a text, but fundamental for its categorization. This process may be partially captured in algorithms provided with the usage of large ontologies or semantic networks (Duch et al., 2008). Our approach is to map words into a network of senses where we use Wordnet (Miller et al., 1993) synsets. First results of creating representations based on synsets are promising—for now we achieved 65% of proper disambiguations (Szymański et al., 2008; Szymański and Duch, 2012).

Evaluation dataset created during the tests formed a kind of initial golden standard dataset for retrieving information in Wikipedia. Currently, it is only for Polish language, but our initiative is to develop it and create datasets similar to TREC data. Creating uniform datasets for evaluation of information retrieval methods in Wikipedia should increase the number of people whose research is involved in processing this huge repository of human knowledge.

TABLE 3 | Results of evaluation WikiClusterSearch for test queries and different numbers of retrieved articles and for different cluster number parameters.

Query	A: 150, C 15; $\varrho = 6.66$			A: 150, C: 30; $\varrho = 3.3$			A: 500, C: 50; $\varrho = 10.0$			A: 500, C: 100; $\varrho = 3.3$		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
kernel	0.67	0.33	0.22	0.85	0.37	0.26	0.74	0.77	0.38	0.82	0.77	0.40
comet	0.69	0.18	0.14	0.81	0.18	0.15	0.86	0.37	0.26	0.92	0.37	0.26
panda	0.75	0.38	0.25	0.89	0.39	0.27	0.73	0.76	0.37	0.78	0.76	0.39
plane	0.82	0.44	0.29	0.94	0.46	0.31	0.89	0.94	0.46	0.94	0.94	0.47
networks	0.63	0.38	0.24	0.61	0.37	0.23	0.92	0.77	0.42	0.95	0.77	0.43
atom	0.60	0.39	0.24	0.72	0.42	0.27	0.88	0.90	0.44	0.93	0.90	0.46
church	0.75	0.41	0.26	0.40	0.40	0.20	0.91	0.82	0.43	0.92	0.82	0.44
pluto	0.90	0.33	0.24	0.99	0.34	0.25	0.84	0.75	0.40	0.85	0.75	0.40
tree	0.53	0.30	0.19	0.82	0.33	0.24	0.73	0.30	0.21	0.91	0.37	0.26
truck	0.79	0.68	0.37	0.85	0.65	0.37	0.81	0.65	0.36	0.64	0.41	0.25
AVG	0.71	0.38	0.25	0.79	0.39	0.26	0.83	0.70	0.37	0.87	0.69	0.37

AUTHOR CONTRIBUTIONS

JS—research concept: algorithms, acquisition of the data. Construction of the experiments (80%). TD—implementation and evaluation of the results (20%).

FUNDING

The work was partially supported by founds of Department of Computer Systems Architecture, Gdańsk University of Technology, Poland.

REFERENCES

Aggarwal, C. C., and Zhai, C. (eds) (2012). "A survey of text clustering algorithms," in *Mining Text Data* (Berlin Heidelberg: Springer), 77–128.

Aljaber, B., Martinez, D., Stokes, N., and Bailey, J. (2011). Improving mesh classification of biomedical articles using citation contexts. *J. Biomed. Inform.* 44, 881–896. doi:10.1016/j.jbi.2011.05.007

Aljaber, B., Stokes, N., Bailey, J., and Pei, J. (2010). Document clustering of scientific texts using citation contexts. *Inf. Retr. Boston* 13, 101–131. doi:10.1007/s10791-009-9108-x

Bennett, C., Li, M., and Ma, B. (2003). Chain letters and evolutionary histories. *Sci. Am.* 288, 76–81. doi:10.1038/scientificamerican0603-76

Bradshaw, S., and Hammond, K. (2002). "Automatically indexing documents: content vs. reference," in *Proceedings of the 7th International Conference on Intelligent User Interfaces* (New York: ACM), 180–181.

Collins, A., and Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407. doi:10.1037/0033-295X.82.6.407

Cvetkovic, D., Doob, M., and Sachs, H. (1995). *Spectra of Graphs – Theory and Applications, III Revised and Enlarged Edition*. Heidelberg–Leipzig: Johan Ambrosius Bart Verlag.

Damashek, M. (1995). Gauging similarity with n-grams: language-independent categorization of text. *Science* 267, 843. doi:10.1126/science.267.5199.843

Deptuła, M., Szymański, J., and Krawczyk, H. (2013). "Interactive information search in text data collections," in *Intelligent Tools for Building a Scientific Information Platform*, eds R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka (Berlin, Heidelberg: Springer), 25–40.

Duch, W. (2009). "Series of information and management sciences," in *8th Int. Conf. on Information and Management Sciences (IMS 2009)* (Kunming-Banna, Yunan, China: California Polytechnic State University), 264–282.

Duch, W., Matykiewicz, P., and Pestian, J. (2008). Neurolinguistic approach to natural language processing with applications to medical text analysis. *Neural Netw.* 21, 1500–1510. doi:10.1016/j.neunet.2008.05.008

Frey, B., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315, 972. doi:10.1126/science.1136800

Grossi, R., and Vitter, J. (2000). "Compressed suffix arrays and suffix trees with applications to text indexing and string matching," in *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, New York, 397–406.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Boston: Elsevier.

Jia, H., Ding, S., Xu, X., and Nie, R. (2014). The latest research progress on spectral clustering. *Neural Comput. Appl.* 24, 1477–1486. doi:10.1007/s00521-013-1439-2

Jiang, Y., Lin, H., Wang, X., and Lu, D. (2011). "A technique for improving the performance of naive Bayes text classification," in *Web Information Systems and Mining*, eds Z. Gong, X. Luo, J. Chen, J. Lei, and F. L. Wang (Berlin Heidelberg: Springer), 196–203.

Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer Verlag.

Kannan, R., and Vetta, A. (2004). On clusterings: good, bad and spectral. *JACM* 51, 497–515. doi:10.1145/990308.990313

Korenius, T., Laurikkala, J., and Juhola, M. (2007). On principal component analysis, cosine and Euclidean measures in information retrieval. *Inf. Sci.* 177, 4893–4905. doi:10.1016/j.ins.2007.05.027

Krebs, C. J. (1999). *Ecological Methodology*, Vol. 2. Menlo Park, CA: Benjamin/Cummings.

Kriegel, H., and Pfeifle, M. (2005). "Density-based clustering of uncertain data," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (New York: ACM), 677.

Mahdi, A. E., and Joorabchi, A. (2010). A citation-based approach to automatic topical indexing of scientific literature. *J. Inf. Sci.* 36, 798–811. doi:10.1177/0165551510388080

Mall, R., Langone, R., and Suykens, J. A. (2013). Kernel spectral clustering for big data networks. *Entropy* 15, 1567–1586. doi:10.3390/e15051567

Manning, C., Raghavan, P., Schütze, H., and Corporation, E. (2008). *Introduction to Information Retrieval*, Vol. 1. Cambridge: Cambridge University Press.

Miller, G. A., Beckitch, R., Fellbaum, C., Gross, D., and Miller, K. (1993). *Introduction to WordNet: An On-line Lexical Database*. Princeton: Cognitive Science Laboratory, Princeton University Press.

Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process Syst.* 2, 849–856.

Ortuño, F. M., Rojas, I., Andrade-Navarro, M. A., and Fontaine, J.-F. (2013). Using cited references to improve the retrieval of related biomedical documents. *BMC Bioinformatics* 14:113. doi:10.1186/1471-2105-14-113

Qian, G., Sural, S., Gu, Y., and Pramanik, S. (2004). "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, New York, 1232–1237.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* 24, 513–523. doi:10.1016/0306-4573(88)90021-0

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Commun. ACM* 18, 613–620. doi:10.1145/361219.361220

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi:10.1109/34.868688

Steinbach, M., Karypis, G., and Kumar, V. (2000). "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, Vol. 400 (Boston: Citeseer), 525–526.

Szymański, J. (2010). "Towards automatic classification of Wikipedia content," in *Springer Lecture Notes in Computer Science, Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL10)*, Berlin Heidelberg, 102–109.

Szymański, J. (2011a). "Categorization of Wikipedia articles with spectral clustering," in *Intelligent Data Engineering and Automated Learning-IDEAL 2011* (Berlin Heidelberg: Springer), 108–115.

Szymański, J. (2011b). "Creating categories for Wikipedia articles using self-organizing maps," in *2011 International Conference on Communications, Computing and Control Applications (CCCA)* (Hammamet, Tunisia: IEEE), 1–5.

Szymański, J. (2013). "Wikipedia articles representation with Matrix'u," in *Distributed Computing and Internet Technology, Volume 7753 of Lecture Notes in Computer Science*, eds C. Hota, P. K. Srimani (Berlin Heidelberg: Springer), 500–510.

Szymański, J. (2014). Comparative analysis of text representation methods using classification. *Cybern. Syst.* 45, 180–199. doi:10.1080/01969722.2014.874828

Szymański, J., and Duch, W. (2010). "Representation of hypertext documents based on terms, links and text compressibility," in *Neural Information Processing. Theory and Algorithms*, eds K. W. Wong, B. S. U. Mendis, and A. Bouzerdoum (Berlin Heidelberg: Springer), 282–289.

Szymański, J., and Duch, W. (2012). "Annotating words using wordnet semantic glosses," in *Neural Information Processing*, eds T. Huang, Z. Zeng, C. Li, and C. S. Leung (Berlin Heidelberg: Springer), 180–187.

Szymański, J., Mizgier, A., Szopiński, M., and Lubomski, P. (2008). Ujednoznacznianie słów przy użyciu słownika WordNet. *Wydawnictwo Naukowe PG TI 2008* 18, 89–195.

Szymański, J., and Wegrzynowicz, K. (2011). "0-step k-means for clustering Wikipedia search results," in *2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)* (Hammamet: IEEE), 253–257.

Vazirani, V. (2013). *Approximation Algorithms*. Springer-Verlag Berlin Heidelberg.

Verma, D., and Meila, M. (2003). *A Comparison of Spectral Clustering Algorithms*. Washington: University of Washington. *Tech. Rep. UW-CSE-03-05-01*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. doi:10.1007/s11222-007-9033-z

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). "Constrained k-means clustering with background knowledge," in *Proceedings of the Eighteenth*

International Conference on Machine Learning, Vol. 577 (San Francisco: Citeseer), 584.

Wang, C., and Duo, C. (2007). An improved density-based dbscan clustering algorithm. *JGXNU* 25, 104.

Wong, S. K. M., Ziarko, W., and Wong, P. N. (1985). "Generalized vector spaces model in information retrieval," in *SIGIR '85* (New York, NY, USA: ACM Press), 18–25.

Wu, W., Xiong, H., and Shekhar, S. (2013). *Clustering and Information Retrieval*, Vol. 11. US: Springer Science & Business Media.

Yang, P., Zhu, Q., and Huang, B. (2011). Spectral clustering with density sensitive similarity function. *Knowl. Based Syst.* 24, 621–628. doi:10.1016/j.knosys.2011.01.009

Zepeda-Mendoza, M. L., and Resendis-Antonio, O. (2013). "Hierarchical agglomerative clustering," in *Encyclopedia of Systems Biology*, eds W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota (New York: Springer), 886–887.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration with the authors, TD and JS, and states that the process nevertheless met the standards of a fair and objective review.