Categorization of Wikipedia articles with spectral clustering

Julian Szymański

Department of Computer Systems Architecture,
Gdańsk University of Technology, Poland,
julian.szymanski@eti.pg.gda.pl

Abstract. The article reports application of clustering algorithms for creating hierarchical groups within Wikipedia articles. We evaluate three spectral clustering algorithms based on datasets constructed with usage of Wikipedia categories. Selected algorithm has been implemented in the system that categorize Wikipedia search results in the fly.

1 Introduction

Documents categorization based on similarity is one of the main goals of automatic knowledge organization. This approach finds many applications especially in information retrieval domain [1] where introducing structure of similarities allows to improve searching for relevant information. In the experiments presented here we evaluate spectral clustering methods used for finding groups of similar documents and organize them into hierarchy. We test three algorithms on datasets constructed from Wikipedia articles. Human made categories of these articles have been used as referential structures, which allows us to construct relevance sets used for evaluation. After evaluation we selected the best clustering method and it has been implemented in practical application that organizes Wikipedia search results into clusters that represent groups of conceptually similar articles instead of their ranked list.

2 Spectral Clustering

One of the important groups of clustering algorithms is spectral clustering. The method is based on cutting the graph of object's similarities using methods of spectral graph theory [2]. In recent years this theory has been strongly developed, especially in direction of graph clustering algorithms where the most well known are: Shi–Malik (2000), Kannan–Vempala–Vetta (2000), Jordan–Ng–Weiss (2002) and Meila–Shi (2000).

If we consider clustering in terms of graph theory we can introduce measure that describe partition of graph nodes into two subsets. If we have weighed graph $G = (V, E)$, it's partition into two sets of nodes (A and B): $A \cap B = \emptyset$ and $A \cup B = V$ can be specified with cutset number [3] defined as:

$$Cut(A, B) = \sum_{u \in A, v \in B} w(u, v). \tag{1}$$

Representing the objects that are to be clustered with graph nodes and the w weights of the graph edges with objects similarities the partitioning problem is reduced to finding optimal cutset. The problem is computationally polynomial [3]. Sometimes cutset gives the results that are different than intuitive nodes partition, so other measures are introduced:

Normalized cut (NCut) (2), that increases its value for clusters that have nodes with small sum of edge weights.

$$NCut(A,B) = \frac{Cut(A,B)}{Vol(A)} + \frac{Cut(A,B)}{Vol(B)} \tag{2}$$

, where

$$Vol(A) = \sum_{u \in A} \sum_{v \in V} w(u,v) \tag{3}$$

Multiway Normalized Cut ($MNCut$) (4) promotes stronger relations between elements in one cluster. It also allows to describe more precisely more separated clusters.

$$MNCut(\Delta) = \sum_{i=1}^{K} (1 - \frac{Cut(C_i, C_i)}{VolC_i}) \tag{4}$$

It is known the spectral clustering methods give high quality partitions [4] and have good computational complexity [3] [5]. There are many variants of spectral algorithms. Mainly they differ in the way of eigenvectors calculation and usage. What is common – they treat source objects as graph nodes and then they are mapped into points in n-dimensional space. This space is constructed using spectral analysis and there is performed essential clustering. We can select three main steps of spectral algorithms ([6]):

1. Preprocessing and data normalization
   At this stage the data are preprocessed into their computational representation. If we are clustering the text documents typically Vector Space Model (VSM) [7] is used. In this representation weighting with Term Frequency and Inverse Document Frequency (TFIDF) [8] allows to calculate similarities between documents. In our experiments we use cosine distance which is known to be the suitable similarity measure for sparse vectors [9].
2. Spectral mapping
   This stage distinguished spectral approach. Using the data from step 1 the typically Laplacian matrix is built and then appropriate number of its eigenvectors is calculated.
3. Clustering
   The objects represented with spectral mapping are divided into two or more sets. Sometimes it is enough to find appropriate cut of the n-element, sorted collection which divide this collection into two clusters. In other methods this step is more complicated and performs partitioning in new representation space (provided by spectral mapping) using standard clustering algorithm eg. k-Means [10].

In our experiments we test three spectral algorithms:

1. Shi–Malik [11] (SM), is realized in following steps:

   (a) Calculate eigenvectors of similarity Laplacian graph.

   (b) Sort elements of the dataset according to second smallest eigenvector value, which is denoted as x_1, x_2, ..., x_n,

   (c) Calculate the partition $\{\{x_1, x_2..., x_i\}, \{x_{i+1}, x_{i+2}, ..., x_n\}\}$ $(1 \leq i \leq n-1)$ having the smallest $NCut$.

   (d) If given partition has $NCut$ value smaller than given a priori value (that means it is better) then this method in each of the divided sets is run again, otherwise the algorithm stops.

2. Kann–Vempala–Vett algorithm [6] (KVV) is

   A heuristic method that finds graph cut which minimizes two parameter quality function, called conductance. Considering partition $(S, \bar{S})$ of graph $G = (V, E)$, where $\bar{S} = V \setminus S$ is a function defined with formula

$$\phi(S) = \frac{Cut(S, \bar{S})}{min(Vol(S), Vol(\bar{S}))} \tag{5}$$

   In comparison to previous one the algorithm instead of Laplacian matrix operates on normalized similarity matrix and uses second biggest (instead second smallest) eigenvector.

   The algorithm goes as follows:

   (a) Normalization performed by dividing each of row elements by the sum of elements in each row.

   (b) For each of clusters C_i created so far, create matrix of similarity $\mathbf{W}_i$ from the nodes the cluster.

   (c) Normalize each matrix $\mathbf{W}_i$ by inserting at the position on diagonal value that complete sum of the elements in this row to 1.

   (d) Calculate second biggest eigenvector of $\mathbf{v}_2^i$ of matrix $\mathbf{W}_i$.

   (e) For each C_i sort its elements according to value of respective coordinate of $\mathbf{v}_2^i$. We denote as x_1^i, x_2^i, ..., $x_{n_i}^i$ sequence of ordered objects form cluster C_i.

   (f) For each C_i find cut in the form of $\left\{\left\{x_1^i, x_2^i..., x_j^i\right\}, \left\{x_{j+1}^i, x_{j+2}^i, ..., x_{n_i}^i\right\}\right\}$ which has smallest conductance.

   (g) Find cluster C_i with the smallest conductance for a given cut and divide it according this cut and replace cluster C_i with two new clusters.

   (h) If given number of clusters has not been reached go to 2.

3. Jordan–Ng–Weiss algorithm [4] (JNW) is

A partitioning algorithm – in one iteration it creates flat clusters, given by a priori K parameter. This parameter denotes also the number of used eigenvectors.
The algorithm performs following steps:
(a) Calculate Laplacian of similarity matrix
(b) Calculate K biggest eigenvectors of Laplacian matrix $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_K$.
(c) Perform spectral mapping from set V (original nodes) to $\mathbb{R}^K$:
$map(i) = [\mathbf{v}_1(i)\ \ \mathbf{v}_2(i)\ \ ...\ \ \mathbf{v}_K(i)]^T$ ($1 \leq i \leq n$).
(d) Perform clustering of the points $map(i)$ in space $\mathbb{R}^K$.
(e) Return result of partitioning of entrance elements into K clusters.
Because JNW is a partitioning approach, the last step is not so simple as in previous algorithms. It requires usage of other clustering method to divide n elements in K-dimensional space. The authors of the algorithm used one of the simplest approaches of k-means.

3 Experiments

In our experiments we compare three algorithms in application to text data. We find Wikipedia[1] dumps very useful to evaluate our approach for documents categorization. These data are easy to preprocess as well as cover wide spectrum of human knowledge thus provide very varied set of documents. What is the most beneficial Wikipedia dumps provide also human made categories that can be used during evaluation process.

3.1 The data

For our experiments we construct eight test data packages. Each package has been constructed with articles from selected Wikipedia super categories and their sub categories retrieved to selected level. The details of each package are shown in Table 1.

3.2 Results

There are many approaches to cluster validation. In our experiments we evaluate results according to external criteria which is known to be harder task than evaluate them according to internal criteria [12]. Our validation we made using standard clustering quality measures (described below) that have been compared to relevance set formed by Wikipedia categories. To make evaluation easier, on each of hierarchy levels we took parameter K (the number of clusters) from referential set.

External cluster validation criteria measure the similarity of the structure of the clusters provided by the algorithm to a priori known structure that is expected to be achieved [13].

Validated cluster structure we denote as $C = \{C_1, ..., C_K\}$, and reference set as $P = \{P_1, ..., P_s\}$. Our source set of objects (articles) is denoted as X, and its cardinality with N. Unordered pairs $\{x_i, x_j\}$ of X elements may fall into four cases:

(a) elements x_i and x_j that belong to the same cluster as well in C as in P,

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

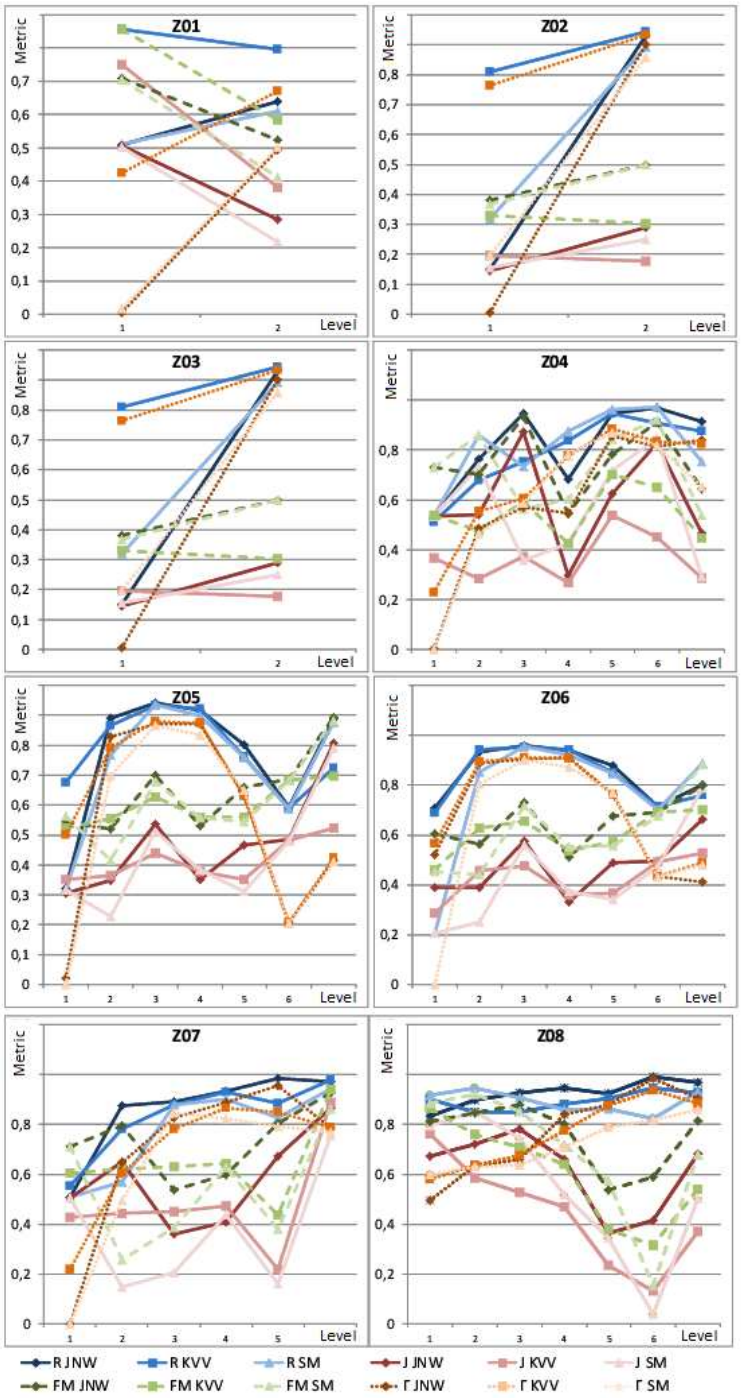

Fig. 1. Results of cluster validation in packages Z01 - Z08.

Table 1. Descriptions: n – number of nodes (in brackets — unisolated), l – non zero elements in neighborhood matrix, N – number of upercategories , k – number of all categories, d – depth of category tree), P – number of categories to Wikipedia root category

n.o.	Name	n	l	N	k	d	P	Comment
1	Z01	575 (575)	67099	2	18	1	3	2 distant categories: Distance_Education i Science_Experiments.
2	Z02	1157 (1156)	323901	5	35	1	3	5 distant categories: Caligraphy, General_Economics, Military_logistics, Evolution, Analytic_number_theory.
3	Z03	3905 (3903)	2260919	8	102	1	3	8 distant categories: Geometric_Topology, Epistemology, Rights, Aztec, Navigation, Clothing_companies, Protests, Biological_Evolution.
4	Z04	3827 (3826)	3195963	2	204	6	4	Two distant categories at the same hierarchy level: Criticism_of_journalism i Corporate_crime.
5	Z05	3647 (3644)	1682361	6	213	6	5	6 distant categories at high level of abstraction: DIY_Culture, Emergence, Military_transportation, Formal_languages, Geology_of_Autralia, Computer-aided_design.
6	Z06	4750 (4747)	2568378	9	289	6	5	9 distant categories at high level of abstraction: DIY_Culture, Emergence, Military_transportation, Formal_languages, Geology_of_Autralia, Computer-aided_design, Special_functions, History_of_ceramics, Musical_thatre_companies.
7	Z07	4701 (4701)	4230139	2	298	6	4	2 neighboring categories at high abstraction level: Computer_law i Prosecution.
8	Z08	5717 (5716)	11288283	4	893	6	4	4 neighboring categories at high abstraction level: Impact_events, Droughts, Volcanic_events, Storm.

(b) elements x_i and x_j that belong to the same cluster in C, but not in P,
(c) elements x_i and x_j that belong to the same cluster in P, but not in C,
(d) elements x_i and x_j that belong to different cluster both in C and in P.

The symbols a, b, c and d denote numbers of elements in respectively cases.

Note they correspond to values in confusion matrix (respectively TP FN FP TN) [14].

Additionally to perform validation two matrices $\mathbf{X}$ i $\mathbf{Y}$ are defined. The first one describes clusters from C, the second from P. Value one at the position (i, j) denotes the pair of elements (x_i, x_j) that belong to different clusters in the structure.

To evaluate our results we use standard clustering quality measures:

- Rand statistics $R = \frac{a+d}{M}$. R value depends on number of objects pairs having their mutual position (same / other cluster) the same in cluster and validation structure. R is in the range $[0, 1]$ and is greater while two compared structures are more similar (1 is when they are identical)
- Jaccard coefficient $J = \frac{a}{a+b+c}$ is similar to R but numerator and denominator are decreased by d value. Similarly to R maximum of Jaccard coefficient is 1 when $b + c = 0$ and it denotes situation when two structures are the same.
- Fowlkes–Mallows index $FM = \frac{a}{\sqrt{(a+b)(a+c)}}$. The FM value grows when a increases and when b and c decrease. The maximum $FM = 1$ when $b = c = 0$.
- Hubert statistics $\Gamma = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i, j)Y(i, j)$. The Γ value grows when the number of elements having same relative position in validated and referential structure increases. $\Gamma_{max} = 1$ when these two structures form N one-element clusters.

In Figure 1 we show metrics values of external validation criteria formed with methods KVV, JNW and SM. Each figure corresponds to the one test package (presented in Table 1), colors denote lines for quality metrics: Rand statistics (R) – blue, Jaccard index (J) – red, Fowlkes–Mallows index (FM) – green and Hubert statistics (Γ) – orange. Different clustering algorithms have been denoted with different line patterns.

4 Discussion

What can be seen from graphs presented in Figure 1 is strong correlation of Rand (R) with Hubert (Γ) statistics as well as Fowlkes–Mallows (FM) with Jaccard index (J). It is caused by high value of a parameter which a denotes number of pairs in one cluster in validated and referential structure. The second parameter that has high influence on metrics is d which denotes a number of objects that were assigned to different clusters in referential and in validation set. Metrics J and FM does not use parameter d, statistics R and Γ involve it in dominator and nominator which cause similarity of these measures.

In some areas we can see differences between metric pairs J and FM as well as R and Γ. Eg. for test package Z04 at level 4 J value for KVV algorithm is low while Γ is high. It is because the structure of the metrics J does not use d parameter, but Γ does. In comparison at 3rd level of hierarchy in this test d value grows (from 3.652.328 to 3.819.940), but a value decreases (from 939.463 to 374.220). Because of the increased d value Γ also increases, while the decrease of a caused the decrease of J value. Similarly there can be substantiated decrease Γ while FM grows (eg. at level 6 of test package Z06). In this case we can see the increase of a and decrease of d. It suggests that on level 6 in referential structure there are fewer categories than at level 5. Indeed number of categories at level 5 is 895 and at level 6 – 503.

The highest values of R and J parameters have been obtained with JNW algorithm. Only one case, when other methods gave better results, is a case described above (level 4 in package Z04). Also FM measure pointed out JNW algorithm as the best. The highest values of Γ in packages Z07 and Z08 we obtain using JNW algorithm, the others have been obtained using KVV.

From the above observations we may conclude the best measure for cluster validation is Rand statistics. It has been less dependent on cluster structure changes, at succeeding hierarchy levels in test packages. It is especially important when a parameter is increasing (growing number of small clusters) in correlation with growing d which causes metric decreases.

In most of the cases the best clusters have been achieved using JNW algorithm, thus we used it in our practical application.

4.1 Practical Application and Future Work

Based on JNW algorithm we have created a prototype system named WikiClusterSearch. It automatically organizes the results of searching Wikipedia for a given keyword. In the system user may specify a searched phrase and the articles containing it are organized into clusters in the fly. It allows to present directions in which user may continue

his or her search. For the efficiency reasons for now only Polish Wikipedia is supported (English one is approximately 5 times bigger). WikiClusterSearch (WCS) has demonstrated the proposed approach can be used to obtain a good quality hierarchy of clusters. The system is available on line under http://swn.eti.pg.gda.pl/UniversalSearch.

In future we plan to develop the JNW approach and use clustering algorithm based on densities [15] instead of k-means. We also plan to implement our system for English Wikipedia what requires to improve its architecture. The long term goal is to join the method of retrieving the information based on clusters of Wikipedia categories with classifier [16] that allows to categorize linear search results returned by search engine into these categories.

We also plan to perform experiments on large scale clustering - it is on the whole Wikipedia. The experiments using well tuned clustering algorithm will allow to improve category system of Wikipedia finding, missing and wrong assignments articles to categories.

ACKNOWLEDGEMENTS

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

References

1. Manning, C., Raghavan, P., Schütze, H., Corporation, E.: Introduction to information retrieval. Volume 1. Cambridge University Press (2008)
2. Cvetkovic, D., Doob, M., Sachs, H.: Spectra of Graphs–Theory and Applications, III revised and enlarged edition, Johan Ambrosius Bart. Verlag, Heidelberg–Leipzig (1995)
3. Vazirani, V.: Algorytmy aproksymacyjne. WNT Warszawa (2005)
4. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems **2** (2002) 849–856
5. Kannan, R., Vetta, A.: On clusterings: Good, bad and spectral. Journal of the ACM (JACM) **51** (2004) 497–515
6. Verma, D., Meila, M.: A comparison of spectral clustering algorithms. University of Washington, Tech. Rep. UW-CSE-03-05-01 (2003)
7. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24** (1988) 513–523
9. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD workshop on text mining. Volume 400., Citeseer (2000) 525–526
10. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning. Volume 577., Citeseer (2001) 584
11. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence **22** (2000) 888–905
12. Eldridge, S., Ashby, D., Bennett, C., Wakelin, M., Feder, G.: Internal and external validity of cluster randomised trials: systematic review of recent trials. Bmj **336** (2008) 876

13. Yeung, K., Haynor, D., Ruzzo, W.: Validating clustering for gene expression data. Bioinformatics **17** (2001) 309
14. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning. Volume 445., Citeseer (1998)
15. Kriegel, H., Pfeifle, M.: Density-based clustering of uncertain data. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM (2005) 677
16. Szymański, J.: Towards automatic classification of wikipedia content. Springer Lecture Notes in Computer Science, Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL10) (2010) 102–109