FINDING SIMILAR DOCUMENTS IN WEB SEARCH RESULTS

Urszula Kużelewska

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Searching the Web is a challenging task. According to the Zamir and Etzioni's definition, Internet is "unorganized, unstructured and decentralized place". Although there are powerful search engines available, the number of indexed web pages exceeds 1 trillion [20] and still grows. Most of the search engines return list of documents from their bases sorted according to their relevance to a search query. Such approach is not the best, because the returned list is very long and may contain documents not related to the query. To increase efficiency of a searching process one may identify groups of similar documents from result list. One of the tools to do it are traditional clustering algorithms. The article presents clustering Web search results directly from a search engine as well as sets created from results for different queries. Documents were grouped using the following methods: EM and XMeans.

Keywords: Web search results clustering, documents similarity, snippets clustering

1. Introduction

Internet is a popular place to share our knowledge or opinion as well as a source of interesting and valuable information. Although the number of web sites is huge, they are useless if they are difficult to find. "The revolution that the Web has brought to information access is not so much due to the availability of information (huge amounts of information has long been available in libraries and elsewhere), but rather the increased efficiency of accessing information, which can make previously impractical tasks practical" [4]. Increased number of information is not related to simultaneous increase in its quality. It requires from the search engines continuous developing and improving standard of generated results.

One of the approaches to this problem is Search Results Clustering (SRC) - techniques of identification groups of similar web sites. According to Weiss [12] the result of SRC are groups of documents, which are organized according to the

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 9, pp. 61-76, 2012.

common theme and described comprehensibly to the human. Such approach does not affect quality or length of a result list, however improves the time of access to relevant information.

The purpose of this paper is to present clustering of Web search results, which have been generated using selected clustering methods with different techniques used in document processing. The main idea of application of traditional clustering algorithms in Web search results domain is not novel, however many new grouping methods have been proposed as well as new procedures in whole process of snippet clustering, which is the reason for testing them in the domains like SRC.

The article is organized as follows: the second section presents short review of methods used in finding similarities in documents from Web search results, the third section describes application of clustering algorithms in SRC domain. The fourth section presents some approaches to the problem of Web search clustering evaluation and the following part contains results of experiments and the last section concludes the paper.

2. Search Results Clustering

Lists of results returned by search engines consist of elements relevant to a query. Each element relates to the particular web page and contains a title, a domain address and a small portion of text from the page (snippet). Search Results Clustering methods work on data preprocessed from titles and/or text of snippets.

The preprocessing of text is as follows: first, all capital letter are replaced by small ones, next, the words without any meaning, such as "is" or "the", are removed, and then the remaining text is stemmed. Stemming is a procedure of extracting constant part of words having different form of inflectional. To give an example, "computer", "computers" or "computerization" could have one common stem - "compute". Words after stemming are determined as terms. The benefits from preprocessing are reduction of the number of words and improvement similarity among elements in final clusters.

One of the steps of Search Results Clustering process is definition of labels describing the generated groups. Depending on the method this part can be performed before or after clustering phase. Typical example solution of this problem is identification of the most frequent words in every group.

Before documents are clustered, they are transformed from their letter representation to numbers in Vector Space Model (VSM) [8]. The numbers relate to the relevance selected words in particular documents. The selected word are called terms.

This process enables application an arbitrary clustering algorithm. The methods of the relevance calculation is described below.

The equation describing a document in Vector Space Model is as follows:

$$D_i = (d_{i1}, d_{i2}, ..., d_{in}) \tag{1}$$

where components d_{ij} refer to the level of description as well as diversification of the individual terms and n is a number of terms selected for representation of documents.

It has been proposed many methods of document description in VSM. One of them is binary representation: if a term from VSM vector is present in the examining document, the relevant component is equal 1. In the other case it is equal 0. More useful are methods based on term or document frequency (TF, DF, TFIDF). One of them, TFIDF, is described as follows:

$$TFIDF(D_i) = TF(t_j, D_i) \cdot IDF(t_j) \tag{2}$$

where component $TF(t_j, D_i)$ refers to the number of occurrences of term t_j in document D_i (see Equation 3) and $IDF(t_j)$ refers to the number of occurrences of this term in all documents (see Equation 4).

$$TF(t_j, D_i) = \begin{cases} 0, & for\, n_{ji} = 0; \\ \frac{n_{ji}}{n_{max}}, & for\, n_{ji} > 0. \end{cases} \tag{3}$$

where n_{ji} denotes a number occurrences of term t_j in document D_i and n_{max} denotes maximal number from occurrences of every term from VSM vector.

$$IDF(t_j) = \log\left(\frac{N}{N_j}\right) \tag{4}$$

where N denotes a number of all documents and N_j - a number of documents containing term t_j.

Documents can be described by all terms present in them, however, to increase time efficiency as well as quality of results terms to VSM vector are selected. It can be used a simple method of selecting the most often terms or one of more complex procedures. The procedures are also based on indices described above, such as TF or IDF.

3. SRC algorithms

Over recent 10 years many important SRC algorithms and systems have been proposed. Despite homogenous contents of clusters, the methods should create compact cluster labels as well as be very effective regarding time. Classical approach to SRC problem applies traditional clustering methods, such as k-means or EM, however there are also systems using different solutions. The most popular are Grouper [14], Carrot [11] and Vivisimo [22]. The algorithms based on traditional clustering are: HKA [5], WISE [1] and ICSE [9].

One of the approaches to document clustering is based on Suffix Tree Clustering (STC) technique, where grouped are phrases instead of individual words. The idea of STC construction has been adapted in Carrot system. The authors created a very extended system with many stemming techniques. They also defined relationship between two main STC parameters: merge threshold and minimum base cluster score, and quality of generated results. A distinguishing feature is dealing with snippets in Polish language.

LINGO was proposed by Osiński [6] in his master thesis and finally became a part of Carrot system. In LINGO was used latent semantic indexing originally adapted in SHOC [15]. In this algorithm the author solve the problem of inadequate labels generation, which occurs in the previous methods, starting the procedure from identification of descriptions of clusters (description comes first). Then documents are assigned to the cluster with the label most matched to their content.

One of the fastest method in traditional clustering domain is k-means. For this reason it is a very popular technique in partitioning document partitioning. Mahdavi and Abolhassani have applied modified k-means method in document clustering domain. They have combined k-means with Harmony Search optimization method to avoid convergence to local optimum and tested the methods on Euclidean as well as cosine similarity/dissimilarity measures.

ICSE (Intelligent Cluster Search Engine) is a system, which also uses k-means algorithm. Documents from a result list are grouped into most relevant, relevant and irrelevant clusters. The clustering method identifies the web pages, which are relevant to the search query in order to increase the relevance rate of search results.

WISE is a meta-search engine, which builds a hierarchical structure of clusters. The clusters contain related web pages expressing one meaning of the query. It uses PoBOC soft clustering algorithm, which is based on graph theory.

4. Evaluation of web search results clustering

Objective assessment of a result of partitioning is a challenging task. Clustering scheme of an internet search list, as well as a clustering result in general, is difficult to evaluate, because it depends on a purpose of the solution and subjective expectations of results' recipients [11]. Considering the reason above, the most useful evaluation is satisfaction of a user searching the information.

However, there are also objective approach to this problem, such as IR (Information Retrieval) indices and criterion merge-then-cluster.

IR indices are composed of precision and recall components. The precision compares a number of documents related to a search query with a number of all documents received from a search engine in answer to the query.

The criterion merge-then-cluster assumes evaluation appropriately prepared set of documents. Original partition is known and compared to generated clustering. The comparison may be performed using traditional measures from clustering evaluation domain [3]. Example of a such index is Rand expressed in the following equation:

$$Rand = \frac{a+d}{a+b+c+d} \tag{5}$$

where a denotes a number of pairs of elements belonging to the same group in both the original as well as the generated solution, b is a number of pairs of elements belonging to different groups in the original, but to the same group in the generated solution, c is a number of pairs of elements belonging to the same group in original, but to different groups in generated solution, d is a number of pairs of elements belonging to different groups in the original as well as in the generated solution.

In experiments presented in this article subjective as well as objective evaluation have been performed.

5. Results of experiments

The system Search Engine used in the following experiments was created at Bialystok University of Technology as a part of a master thesis "An intelligent search engine using clustering methods to optimize search results" [10]. The results were generated by Bing [16] search engine and the clustering algorithms were taken from Weka system [13]. The system allows data entry as XML file, as well, which allows objective evaluation of quality of results. A browser window of the system is presented in Figure 1.

In the system the step of clustering may be realised by one of the algorithms: EM [2] and XMeans [7]. EM clusters elements basing on probability of their membership to each group. In the system, it is possible to run EM without giving the information about a number of clusters. XMeans is an extension of k-means method. One of improvements is automatic calculation of a number of groups. However a user is required to give this value, the final result may contain a different, more optimal, number of clusters.

The experiments consist of 3 parts: on-line clustering data from Bing Web Service, clustering data from file containing different as well as similar description of elements between clusters (created from results of various queries) and clustering benchmark data from Credo repository [19].

5.1 Experiments with data clustering from a search engine

In this experiment, the query "*Java*" was entered in Bing search engine and returned 100 results. The results were clustered using 4 algorithms: Lingo, STC [18], XMeans and EM [10]. In case of methods requiring input parameters (XMeans and in some cases EM), their values were adjusted to obtain the smallest number of clusters as well as equal clusters' size. EM was started with the following values of parameters: unknown number of clusters, length of description vector: 25, documents description method: TFIDF. XMeans was started with the numbers of clusters equal 20, length of description vector: 45 and documents description method: TFIDF.

Tables 1 and 2 show a brief summary of the results generated by respectively Lingo and STC algorithms and XMeans and EM methods, which allows subjective evaluation and comparison of generated results. The tables contain only larger clusters (of size greater than 3 elements).

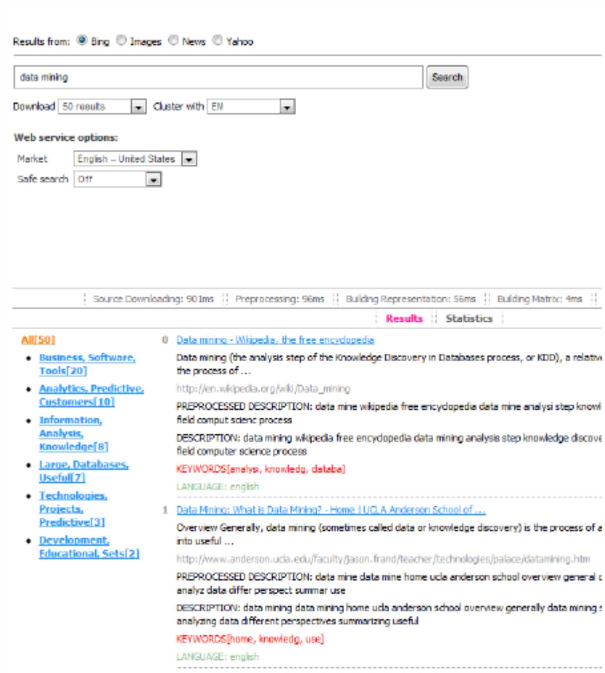
Lingo algorithm split the result list in many groups containing small number (10 and less) of elements, whereas STC, unfortunately, has identified one large group composed of 88 items and several smaller clusters. XMeans and EM methods generated several groups ranging in sizes from several to 40 items. Labels of the groups concerned domains: computer technology and programming and were phrases (in case of Carrot system) or consisted of one word (in case of SearchEngine program). The greatest group of STC method had label *Java*, which is undesirable, because it equals the input query. However, interesting labels were created in case of XMeans algorithm: *Sun* and *Oracle*, concerning companies connected with Java programming language.

Table 1. Clustering results of *Java* query generated by Lingo and STC methods

Method	Group label	Number of results	Method	Group label	Number of results
Lingo	Java Tutorials	10	STC	Java	88
	Java Coffee	8		Programming	16
	Java.net	8		Download	12
	Java Technology	7		Source	12
	Downloads	6		Software	11
	Java Community	6		Tutorials	11
	Java Language	6		Coffee	9
	Learn Java	6		Java Programming	8
	Open Source	5		Java.net	8
	Source Code	5		Developers	8
	Standard Edition	5		Standard Edition	7
	Tutorial	5		Java Software	6
	Browser	4		Open Source	5
	Guide	4		Virtual Machine	5
	Java Programming Resources	4	-	-	-
			-	-	-
	Virtual Machine	4	-	-	-

Table 2. Clustering results of *Java* query generated by XMeans and EM methods

Method	Group label	Number of results	Method	Group label	Number of results
XMeans	Development	23	EM	Programming	40
	Programming	18		Developers	21
	Information	9		Tutorials	19
	Code	8		Download	14
	Software	7		Software	5
	Sun	7		-	-
	Oracle	6	-	-	-
	Learn	5	-	-	-



Results from: ● Bing ○ Images ○ News ○ Yahoo

data mining

Download 50 results Cluster with EN

Web service options:

Market English – United States

Safe search Off

Source Downloading: 901ms ¦¦ Preprocessing: 96ms ¦¦ Building Representation: 56ms ¦¦ Building Matrix: 4ms

All[50]

- Business, Software, Tools[20]
- Analytics, Predictive, Customers[10]
- Information, Analysis, Knowledge[8]
- Large, Databases, Useful[7]
- Technologies, Projects, Predictive[3]
- Development, Educational, Sets[2]

0 Data mining - Wikipedia, the free encyclopedia

Data mining (the analysis step of the Knowledge Discovery in Databases process, or KDD), a relativ
the process of ...

http://en.wikipedia.org/wiki/Data_mining

PREPROCESSED DESCRIPTION: data mine wikipedia free encyclopedia data mine analysi step knowl
field comput scienc process

DESCRIPTION: data mining wikipedia free encyclopedia data mining analysis step knowledge discove
field computer science process

KEYWORDS[analysi, knowledg, databa]

LANGUAGE: english

1 Data Mining: What is Data Mining? - Home | UCLA Anderson School of ...

Overview Generally, data mining (sometimes called data or knowledge discovery) is the process of a
into useful ...

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

PREPROCESSED DESCRIPTION: data mine data mine home ucla anderson school overview general c
analyz data differ perspect summar use

DESCRIPTION: data mining data mining home ucla anderson school overview generally data mining s
analyzing data different perspectives summarizing useful

KEYWORDS[home, knowledg, use]

LANGUAGE: english

Fig. 1. Example window of Search Engine system

Tables 3 and 4 contains selected pages from the result list returned by Bing search engine and group labels to which they have been assigned by 4 examined clustering methods.

The results are slightly different depending on the algorithm, however each of clustering scheme is correct regarding content of clusters (excluding STC). The pages: "*The Java^{TM} Tutorials*" and "*Learn Java – Tutorials, Tips, Help...*" were placed depending on the algorithm: in the group *Java Tutorials* (Lingo), *Tutorials* (EM) and in two groups *Programming* and *Learn* (XMeans). One of the mentioned pages was assigned by Lingo to two different groups (*Java Tutorials* and *Java Language*). The page from Wikipedia ("*Java (programming language) – Wikipedia, ...*") was clustered as *Java Language* (Lingo), *Sun* (XMeans) or *Programming* (EM) and the page "*Java Programming Resources – Java, Java, and ...*" was assigned to the cluster *Java Tutorials* (Lingo) and *Programming* (XMeans and EM). The remaining two pages concerning technology was grouped by Lingo to the cluster *Java Technology* and to the cluster *Developers* (XMeans and EM),

Table 3. Part of clustering results of *Java* query generated by Lingo and STC methods (some URLs have been shortened)

Page title & Page URL	Group label	
	Lingo	STC
The Java$^{\text{TM}}$ Tutorials http://download.oracle.com/javase/tutorial/	Java Tutorials Java Language	Java
Learn Java – Tutorials, Tips, Help ... http://java.about.com/	Java Tutorials	Java
Java (programming language) - Wikipedia, ... http://en.wikipedia.org/.../Java_(programming...	Java Language	Java
Java Programming Resources – Java, Java, and ... http://www.apl.jhu.edu/ hall/java/	Java Tutorials	Java
Oracle Technology Network for Java Developers http://www.oracle.com/.../java/index.html	Java Technology	Java
Nokia Developer - Java http://www.developer.nokia.com/Develop/Java/	Java Technology	Java

Table 4. Part of clustering results of *Java* query generated by XMeans and EM methods (some URLs have been shortened)

Page title & Page URL	Group label	
	XMeans	EM
The Java$^{\text{TM}}$ Tutorials http://download.oracle.com/javase/tutorial/	Programming	Tutorials
Learn Java – Tutorials, Tips, Help ... http://java.about.com/	Learn	Tutorials
Java (programming language) - Wikipedia, ... http://en.wikipedia.org/.../Java_(programming...	Sun	Programming
Java Programming Resources – Java, Java, and ... http://www.apl.jhu.edu/ hall/java/	Programming	Programming
Oracle Technology Network for Java Developers http://www.oracle.com/.../java/index.html	Oracle	Developers
Nokia Developer - Java http://www.developer.nokia.com/Develop/Java/	Developers	Developers

however, only XMeans assigned one page connected with Oracle company to *Oracle* group. Unfortunately, STC algorithm placed all of the pages in one group *Java*.

5.2 Experiments with small datasets

In this experiment datasets were created from Google [17] search results of queries concerning cities names. It has been selected snippets from two initial pages of results. The queries were composed to examine abilities of the methods (XMeans and EM) to separate completely different as well as slightly similar clusters.

The first case of data, $2cities$, contains results of *Warsaw* nad *New York* queries. Numbers of snippets in each list were 18, 20 and 20 respectively (see Table 5). In the second case - $3cities$ - results of query *London* have been added to the previous file. Finally, in the last data set - $3cities\&airport$ - snippets from results of *Warsaw airport* query have been joined.

Despite of subjective evaluation of clusterings (see summary of selected results in Tables 6 and 7), it was performed objective assessment - calculation Rand index (see Table 8).

Table 5. Components (queries) of $2cities$, $3cities$ and $2cities\&airport$ data

Data set	Query	Number of results
2cities	Warsaw	18
2cities	New York	20
3cities	Warsaw	18
3cities	New York	20
3cities	London	20
3cities&airport	Warsaw	18
3cities&airport	New York	20
3cities&airport	London	20
3cities&airport	Warsaw airport	10

Table 6. Results of clustering into 3 groups of $3cities\&airport$ data generated by EM method

Group label	Number of elements	Original components
New	16	16(New York)
London	24	2(Warsaw), 3(New York), 19(London)
Warsaw	29	17(Warsaw), 2(London), 10(Airport)

Table 7. Results of clustering into 4 groups of 3*cities&airport* data generated by EM method

Group label	Number of elements	Original components
New	20	20(New York)
London	17	1(Warsaw), 2(Airport), 14(London)
Warsaw	26	18(Warsaw), 6(London), 2(Airport)
Airport	6	6(Airport)

Table 8. Results of clustering (Rand coefficient) of 2*cities*, 3*cities* and 2*cities&airport* data

Data set	Method name	Number of groups	Rand
2cities	EM	-	0.55
2cities	EM	2	0.95
3cities	EM	-	0.75
3cities	EM	3	0.69
3cities	XMeans	3	0.85
3cities&airport	EM	3	0.81
3cities&airport	EM	4	0.86

An interesting experiment was one, when data contained 3 main groups (*New York*, *Warsaw*, *London*) and one another - *Warsaw airport*, which might be a subgroup of *Warsaw* cluster. Tables 6 and 7 show results of clustering the data, when a number of groups was set to 3 and 4 respectively.

In the first case all the results from query *Warsaw airport* is clustered to *Warsaw* group (forming a subgroup), whereas in the following experiment, in which the number of groups is equal to the number of result lists - most of them is separated into additional cluster.

5.3 Experiments with large dataset

The dataset - *ambiguous* - is taken from Credo repository. The topics of the queries were selected from the ambiguous Wikipedia list. Elements of the list contain the word "disambiguation" in the titles, e. g. "Aida" is a title of opera by Giuseppe Verdi, as well as "a set of defined interfaces and formats for representing common data analysis objects, primarily used by researchers in high-energy particle physics" [21]. In the experiments it has been used 100 results concerning 10 main topis, which were clustered using EM and XMeans methods.

In both cases of algorithms, a number of clusters was given: 10 in EM and ranging from 10 to 20 in XMeans. The other parameters were as follows: documents description method: TFIDF and length of description vector was ranging from 10 to 20. Selected results are shown in Tables 9, 10 and 11.

Table 9. Clustering results of *ambiguous* data generated by EM method

Group label	Number of elements	Original components
B-52	98	98 (B-52)
Bronx	97	97 (Bronx)
Cube	91	91 (Cube)
Aida	92	92 (Aida)
Eos	92	92 (Eos)
Camel	98	98 (Camel)
Beagle	97	97 (Beagle)
Sea	1	1 (Aida)
Coral	238	7 (Aida), 3 (Bronx) 2 (B-52), 4 (Cain) 3 (Beagle), 2 (Camel) 100 (Coral Sea), 9 (Cube) 8 (Eos), 100 (Excalibur)
Cain	96	96 (Cain)

Tables show, that the identified groups are homogenous in most of cases. The groups generated by EM method (see Table 9) contain more than 90 of 100 elements from the original partition. The only exception are groups: *Sea*, which is composed of 1 element and *Coral* containing mostly 2 original groups: *Coral Sea* and *Excalibur*.

Table 10 presents a result created by XMeans method, when a number of groups was set on 10. The algorithm generated 8 groups: 5 large homogenous (70-90 elements) clusters, 2 small, but composed of only one original partition. Unfortunately, there is also the greatest cluster, which contains 5 original groups.

In the following part of experiment, the number of groups given to XMeans has been increased to 20. The method has generated 16 clusters, which sizes were ranging from 2 to 94. Contrary to the previous part, in this the formed groups (except for 2) were homogenous.

In this experiment the results were also evaluated by Rand index. Its value for EM result (see Table 9) was 0.956, whereas for XMeans partitioning were 0.77 (see Table 10) and 0.966 (see Table 11). The high values (around 1) indicate great compatibility the generated results with original groups.

Table 10. Clustering results (8 groups) of *ambiguous* data generated by XMeans method

Group label	Number of elements	Original components
Bronx	98	98 (Bronx)
Cube	88	88 (Cube)
Aida	76	76 (Aida)
Camel	89	89 (Camel)
Music	11	11 (Camel)
Sea	17	17 (Aida)
Coral	521	7 (Aida), 2 (Bronx) 100 (B-52), 1 (Cain) 100 (Beagle), 100 (Eos) 100 (Coral Sea), 12 (Cube) 100 (Excalibur)
Cain	99	99 (Cain)

6. Conclusions

The purpose of this paper was to present possibilities of application of clustering methods in grouping Web search results. Two algorithms were selected and used in the experiments - EM and XMeans. The experiments were divided into 3 parts: clustering on-line snippets from a search engine, verification of ability to discover different as well as similar clusters and clustering large data containing ambiguous search results. The results were evaluated by Rand index or compared to partitionings from Lingo and STC systems.

There are many improvements to make, however the presented results show great usefulness of traditional clustering methods in the domain of SRC. In the first, on-line experiment, the created clusters consisted of similar snippets, which were described by adequate labels. Moreover, the groups were not fragmented and the labels were diversified, as well.

In the remaining experiments, in most cases, original clusters were properly identified by the examined methods: EM and XMeans. Generated clusters in many results were homogenous and Rand coefficient was about 0.8. It is particularly evident in case of ambiguous data. Labels of clusters were relevant to its content, however in future it is desirable to describe them by phrases. It is worth recalling the fact of identification of a subgroup in the second experiment.

It may be concluded, that as long as clustering algorithms are proposed they should be checked in SRC domain. Particularly interesting are methods generating

Table 11. Clustering results (16 groups) of *ambiguous* data generated by XMeans method

Group label	Number of elements	Original components
Aida	87	87 (Aida)
Cube	81	81 (Cube)
Camel	61	61 (Camel)
Eos	75	75 (Eos)
Bronx	87	87 (Bronx)
Information	33	33 (Camel)
Amp	2	2 (Cube)
B-52	85	85 (B-52)
Coral	84	84 (Coral Sea)
Sea	17	15 (Coral Sea), 1 (Aida) 1 (Eos)
Musicals	72	8 (Excalibur), 10 (Aida) 9 (B-52), 8 (Bronx) 4 (Beagle), 3 (Camel) 11 (Cain), 1 (Coral Sea) 7 (Eos), 11 (Cube)
Cain	86	86 (Cain)
Beagle	94	94 (Beagle)
Excalibur	89	89 (Excalibur)
Reviews	30	3 (Excalibur), 2 (Aida) 1 (B-52), 3 (Bronx) 2 (Cain), 5 (Cube) 14 (Eos)
Photo	17	5 (B-52), 2 (Bronx) 1 (Cain), 1 (Cube) 2 (Beagle), 3 (Eos) 3 (Camel)

compact and separable groups as well as able to identify hierarchical relationships in clustering results.

References

[1] R. Campos, G. Dias, C. Nunes, WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques, Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 301-304

[2] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, 39, 1977, pp. 1–38

[3] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems, 17(2/3), 2001, pp. 107–145

[4] S. Lawrence, C. L. Giles, Accessibility and Distribution of Information on the Web, Nature (400), 1999, pp. 107-109

[5] M. Mahdavi, H. Abolhassani, Harmony K-means algorithm for document clustering, Data Mining and Knowledge Discovery(18), 2009, pp. 370–391

[6] S. Osiński, An algorithm for clustering of web search results, Master Thesis, Poznan University of Technology, 2003

[7] D. Pelleg, A. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, Proceedings of International Conference on Machine Learning, 2000, pp. 727-734

[8] G. Salton, A Vector Space Model for Automatic Indexing, Communications of the ACM, 18(11), 1975, pp. 613-620

[9] M. Sathya, J. Jayanthi, N. Basker, Link Based K-Means Clustering Algorithm for Information Retrieval, Proceedings of IEEE-International Conference on Recent Trends in Information Technology, 2011, pp. 1111-1115

[10] W. Rakowski, An intelligent search engine using clustering methods to optimize search results, Master Thesis (in Polish), Bialystok University of Technology, 2011

[11] D. Weiss, A Clustering Interface for Web Search Results in Polish and English, Master Thesis, Poznan University of Technology, 2001

[12] D. Weiss, The search for meaning in a haystack, Seminar of Institute of Linguistics, Polish Academy of Science (in Polish), 2003

[13] I.H. Witten, E. Frank, M.A. Hall, Weka: data mining software in Java, `[http://www.cs.waikato.ac.nz/ml/weka/]` (26.06.2012)

[14] O. Zamir, O. Etzioni, Grouper: A Dynamic Clustering Interface to Web Search Results, WWW Computer Networks 31(11-16), 1999, pp. 1361-1374

[15] D. Zhang, Y.Dong, Semantic, Hierarchical, Online Clustering of Web Search Results, APWeb'2004, 2004, pp. 69-78

[16] Bing Search Engine, [http://www.bing.com] (10.09.2011)

[17] Google Search Engine, [http://www.google.pl] (15.10.2012)

[18] Carrot2 Clustering Engine, [http://search.carrot2.org/stable/search]

[19] Web search results datasets, [http://credo.fub.it/] (26.06.2012)

[20] Google Blog, [http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html], (21.06.2012)

[21] Wikipedia description of Aida expression, [http://en.wikipedia.org/wiki/Aida_%28disambiguation%29]

[22] Vivisimo company, [http://vivisimo.com] (10.09.2011)

IDENTYFIKOWANIE DOKUMENTÓW PODOBNYCH W WYNIKACH WYSZUKIWANIA W SIECI WWW

Streszczenie: Przeszukiwanie sieci WWW jest niezmiernie trudnym zadaniem. Według Zamira i Etzioniego Internet to "miejsce bez struktury, niezorganizowane i zdecentralizowane". Chociaż istnieją potężne narzędzia w postaci wyszukiwarek internetowych, ich użycie staje się z czasem trudniejsze, gdyż ilość zaindeksowanych stron internetowych przekracza 1 bln [20] i nadal rośnie. Większość wyszukiwarek generuje wyniki posortowane według ich zgodności z treścią zapytania w postaci bardzo długich list. Takie podejście nie jest najlepszym rozwiązaniem z powodu rozmiaru list oraz zawierania w nich dokumentów nie związanych z zapytaniem. W celu zwiększenia efektywności przeszukiwania Internetu można zastosować grupowanie podobnych dokumentów z generowanej przez wyszukiwarki listy wyników. Jednym z takich narzędzi są tradycyjne algorytmy grupujące. W artykule przedstawiono wyniki grupowania dokumentów bezpośrednio z listy zwróconej przez wyszukiwarkę oraz zbiorów dokumentów utworzonych z wyników wyszukiwania dla kilku zapytań. Wykorzystano następujące metody grupujące: EM i XMeans.

Słowa kluczowe: grupowanie wyników wyszukiwania, podobieństwo dokumentów, grupowanie snippetów

Artykuł zrealizowano w ramach pracy badawczej nr S/WI/5/08.