A.A. Krizhanovsky

EXPERIMENTS ON CORPUS INDEX GENERATED FROM WIKIPEDIA[1]

1. Introduction

In the USA, the 2007 nationwide survey found that more than a third of adult Internet users (36%) consulted the online encyclopedia Wikipedia[2]. The popularity of encyclopedia is probably best explained by the sheer amount of material on the site, the wide coverage of topics and the freshness of data. Wikipedia (WP) continues to gain popularity among the broad masses because it has a high rank assigned by search engines. E.g., in March 17, 2007, over 70% of the visits to Wikipedia came from search engines, according to Hitwise data[2]. More over, the search system Koru analyses Wikipedia links to expand query terms[3].

There are two kinds of data in Wikipedia: text and links (internal, external, interwiki, categories). Accordingly, three types of search algorithms[4] could be applied to the Wikipedia data:

[1] This work is supported in part by grant # 08-07-00264 of the Russian Foundation for Basic Research and projects supported by the Russian Academy of Sciences (RAS) # 14.2.35 and # 1.9.Full version of this paper is available at **http://arxiv.org/abs/0808.1753**

[2] *Rainie L., Tancer B.* Wikipedia users // Reports: Online Activities & Pursuits. 2007 // URL: **http://www.pewinternet.org/pdfs/PIP_Wikipedia07.pdf**

[3] *Milne D., Witten I.H., Nichols D.M.* A knowledge-based search engine powered by Wikipedia // In Proc. of the ACM Conference on Information and Knowledge Management (CIKM'2007). Portugal, Lisbon, 2007 // URL: **http://www.cs.waikato.ac.nz/~dnk2/publications/ cikm07.pdf**

[4] We are interested in algorithms that either search for documents by keywords, or search for documents that are similar to the original one.

1) Link analysis algorithms that, in their turn, may be classified into two categories:

- links are defined explicitly by hyperlinks (HITS[1], PageRank[2], ArcRank[3], Green[4], WLVM[5]);
- links are built automatically (Similarity Flooding[6], automatic synonym extraction in a dictionary[7]);

[1] *Kleinberg J.*, Authoritative sources in a hyperlinked environment // Journal of the ACM. 1999. № 5 (46). P. 604–632 // URL: **http://www.cs.cornell.edu/home/kleinber**

[2] *Brin S., Page L.* The anatomy of a large-scale hypertextual Web search engine. 1998 // URL: **http://www-db.stanford.edu/~backrub/google.html**

[3] Survey of text mining: clustering, classification, and retrieval / M. Berry (Ed.). Springer-Verlag, New York, 2003. 244 pp.

[4] *Ollivier Y., Senellart P.* Finding related pages using Green measures: an illustration with Wikipedia // Association for the Advancement of Artificial Intelligence.Vancouver, Canada, 2007 // URL: **http://pierre.senellart.com/publications/ollivier2006finding.pdf**

[5] *Milne D.* Computing Semantic Relatedness using Wikipedia Link Structure // New Zealand Computer Science Research Student Conference (NZCSRSC'2007). Hamilton, New Zealand // URL: **http://www.cs.waikato.ac.nz/~dnk2/publications/nzcsrsc07.pdf**

[6] *Melnik S., Garcia-Molina H., Rahm E.* Similarity flooding: a Versatile graph matching algorithm and its application to schema matching // 18th ICDE. San Jose CA, 2002 // URL: **http://research.microsoft.com/~melnik/publications.html**

[7] *Blondel V., Senellart P.* Automatic extraction of synonyms in a dictionary // Proceedings of the SIAM Workshop on Text Mining. Arlington (Texas, USA), 2002 // URL: **http://www.inma.ucl.ac.be/~blondel/publications/areas.html**

2) Statistical text analysis (ESA[1], the similarity of short texts[2], constructing contextually similar words[3], the self-organizing map[4];

3) Text and link analysis[5].

The earlier developed adapted HITS algorithm (AHITS) [6] searches for related terms by analysing Wikipedia internal links. There are many algorithms for searching related terms in Wikipedia, which

[1] *Gabrilovich E., Markovitch S.* Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis // Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India, January, 2007 // URL: **http://www.cs.technion.ac.il/~gabr/papers/ijcai-2007-sim.pdf**

[2] *Sahami M., Heilman T.D.* A web-based kernel function for measuring the similarity of short text snippets // In Proceedings of the 15th International World Wide Web Conference (WWW), 2006 // URL: **http://robotics.stanford.edu/users/sahami/papers-dir/www2006.pdf**

[3] *Pantel P., Lin D.* Word-for-word glossing with contextually similar words. In Proceedings of ANLP-NAACL 2000. Seattle, Washington, May, 2000. P. 75–85 // URL: **http://www.cs.ualberta.ca/~lindek/papers.htm**

[4] *Lee C.H., Yang H.C.* Acquisition of Web semantics based on a multilingual text mining technique // In 2nd Workshop at ECML/PKDD-2002. Finland, Helsinki, 20 August, 2002. P. 61–78 // URL: **http://km.aifb.uni-karlsruhe.de/ws/semwebmine2002/papers/application/lee_yang.pdf**

[5] *Bharat K., Henzinger M.* Improved algorithms for topic distillation in a hyperlinked environment // Proc. 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), 1998. P. 104–111 // URL: **ftp://ftp.digital.com/pub/DEC/SRC/publications/monika/sigir98.pdf**; *Maguitman A. G., Menczer F., Roinestad H., Vespignani A.* Algorithmic Detection of Semantic Similarity. 2005 // URL: **http://www2005.org/cdrom/contents.htm**

[6] *Krizhanovsky A.* Synonym search in Wikipedia: Synarcher // 11-th International Conference «Speech and Computer» SPECOM'2006. Russia, St. Petersburg, June 25–29, 2006. P. 474–477 // URL: **http://arxiv.org/abs/cs/0606097**

can do without full text search[1] (Table 3, p. 8). However, experimental comparison of algorithms[2] shows that the best results were obtained with the statistical text analysis algorithm ESA.

This induce us to create the publicly available index database of Wikipedia (further referred to as WikIDF[3]) and tools for database creation, which, as a whole, provides a full text search in the encyclopedia in particular, and in MediaWiki-based[4] wiki sites in general. Wikitext is text in a markup language that offers a simplified alternative to HTML[5]. Markup tags are useless for keyword search, and hence the wikitext should be converted to a text in natural language at the preliminary stage of indexing.

[1] *Krizhanovsky A.* Evaluation experiments on related terms search in Wikipedia: Information Content and Adapted HITS (In Russian) // Proc. of the Wiki-Conference 2007, Russia, St. Petersburg, October 27–28 // URL: **http://arxiv.org/abs/0710.0169**

[2] *Gabrilovich E., Markovitch S.* Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis // Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India, January, 2007 // URL: **http://www.cs.technion.ac.il/~gabr/papers/ijcai-2007-sim.pdf**

[3] The WikIDF abbreviation reflects the fact that the generated corpus index is suitable for TF-IDF weighting scheme.

[4] MediaWiki is a web-based wiki software application used by many wiki sites, e.g. Wikipedia.

[5] See URL: **http://en.wikipedia.org/wiki/Wikitext**

The wiki indexing system integrates the programs: GATE[1], Lemmatizer[2] and Synarcher[3]. The architecture of WikIDF is described in the paper[4]. WikIDF is a console application (part of Synarcher[5] program), which depends on the RuPOSTagger[6] program. WikIDF is bundled with Synarcher.

The developed software (the database and the indexing system) will allow scholars to analyse the obtained Wikipedia index database, and programmers to use this system as a part of their search engine in order to retrieve information from wiki sites.

2. Experiments

The developed software for indexing wiki-texts enabled to create an index databases of Simple English Wikipedia[7] (further, denote SEW) and Russian Wikipedia[8] (RW) and to carry out experiments.

[1] *Cunningham H., Maynard D., Bontcheva K., Tablan V., Ursu C., Dimitrov M., Dowman M., Aswani N., Roberts I.* Developing language processing components with GATE (user's guide), Technical report, University of Sheffield, U.K., 2005 // URL: **http://www.gate.ac.uk**

[2] *Sokirko A.* A short description of Dialing Project. 2001 // URL: **http://aot.ru/docs/sokirko/sokirko-candid-eng.html**

[3] *Krizhanovsky A.* Synonym search in Wikipedia: Synarcher…

[4] *Smirnov A., Krizhanovsky A.* Information filtering based on wiki index database // Proceedings of the 8th International FLINS Conference on Computational Intelligence in Decision and Control. Spain, Madrid, September 21–24, 2008 // URL: **http://arxiv.org/abs/0804.2354**

[5] See **http://synarcher.sourceforge.net**

[6] See more information about Lemmatizer and RussianPOSTagger at **http://rupostagger.sourceforge.net**

[7] Most frequent 1000 words found in English Simple Wikipedia (14 Feb 2008) are listed with frequencies, see **http://simple.wiktionary.org/wiki/User:AKA_MBG/English_Simple_Wikipedia_20080214_freq_wordlist**

[8] Most frequent 1000 words found in Russian Wikipedia (20 Feb 2008), see **http://ru.wiktionary.org/wiki/Конкорданс:Русскоязычная_Википедия/20080220**

The statistical data of the source / result databases and the parsing process are presented in Table 1.

In two columns («RW / SEW 07» and «RW / SEW 08») the values of the RW parameters (at 20/09/2007 and 20/02/2008) divided by the SEW parameters (at 09/09/2007 and 14/02/2008) in 2007 and 2008 years, respectively, are presented. The parameters that characterize the Russian Wikipedia are the large quantity of lexemes (1,43 M[1]) and the total number of words in the corpus (32,93 M).

The size of Russia Wikipedia is an order of magnitude higher than Simple English one (column «RW/SEW 08»): the number of articles is 9,5 times greater, the number of lexemes is 9,6 times, the number of total words is 14,4 times.

The values in the next two columns («SEW 08/07%» and «RW 08/07%») show how much the sizes of English and Russian corpuses (in comparison with itself) are increased during the five months from September 2007 to February 2008.

The last column (SEW↑ /RW↑) shows how much the rate of enlargement of the English corpus in comparison with Russian one (the division of values of the previous two columns), namely, by 12% faster creation of new articles, and by 6% faster enriching the lexicon of Wikipedia in Simple English.

The following hardware, software and versions of the two main programs were used in experiments presented in Table 1: OS Debian 4.0 etch, Linux kernel 2.6.22.4, the AMD processor 2,6 GHz, 1 GB RAM, Java SE 1.6.0_03, MySQL 5.0.51a-3.

[1] *M* symbol denotes million, see **http://en.wikipedia.org/wiki/SI_prefix**

Table 1. The statistical data of Wikipedias and created index databases

Wikipedias	**Simple English (SEW 08)**	**Russian (RW 08)**	**RW / SEW 07**	**RW / SEW 08**	**SEW 08/07 %**	**RW 08/07 %**	**SEW↑ / RW↑ %**
Wikipedia Database							
Database dump, timestamp	14/02/2008	20/02/2008	–	–	–	–	–
Database dump, size, MB	21.11	240.38	15.9	14.4	40	26	10
Articles, k.	25.22	239.29	10.7	**9.5**	31	17	**12**
Index Database of Wikipedia							
Lexemes in the corpus, M	0.149	**1.43**	10.2	**9.6**	23	16	**6**
Lexeme-page (<=1000 for 1 lexeme), M	1.65	15.71	10.1	9.5	24	16	6
Words in the corpus, M	2.28	**32.93**	15.1	**14.4**	29	23	5
Size of archived file of index DB dump, MB	7.15	77.5	11.5	10.8	25	17	6

In Table 1 «Lexeme-page» shows the number of relations «lexeme-page» extracted from the corpus. It could be stored (to the index DB) no more than 1000 relations for one lexeme. The number 1000 is one of the input parameters of the indexing program FLINS.

3. Conclusions

With the fantastic growth of Internet usage, information search in documents of a special type called a «wiki page» that is written using a simple markup language, has become an important problem.

The wiki texts indexing application was developed. This paper describes the properties and comparison of corpus indexes generated from the text corpus of Russian Wikipedia (RW) and Simple English Wikipedia (SEW)[1].

The size of RW is by order of magnitude higher than SEW (number of words, lexemes), though the growth rate of number of pages in SEW was found to be 12% higher than in Russian, and the rate of acquisition of new words in SEW lexicon was 6% higher during a period of five months (from September 2007 to February 2008).

[1] The created index DB for Russian and Simple English WP are available at: **http://rupostagger.sourceforge.net**, see packages *idfruwiki* and *idfsimplewiki*.