# Synonym Search in Wikipedia: Synarcher

*Andrew A. Krizhanovsky*

Computer-Aided Integrated Systems Laboratory
St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
St.-Petersburg, Russia
aka@mail.iias.spb.su

## Abstract

The program Synarcher for synonym (and related terms) search in the text corpus of special structure (Wikipedia) was developed. The results of the search are presented in the form of graph. It is possible to explore the graph and search for graph elements interactively. Adapted HITS algorithm for synonym search, program architecture, and program work evaluation with test examples are presented in the paper. The proposed algorithm can be applied to a query expansion by synonyms (in a search engine) and a synonym dictionary forming.

## 1. Introduction

Growing number of documents, appearance of new auxiliary structures and metadata linked to documents enable one to adapt known algorithms and to propose new ones for more precise search. The similarity search problem includes search of similar text documents, semantically related words, and graph similar vertices. There are enough algorithms for solving this problem, e.g., the "hypertext induced topic selection"(HITS) algorithm proposed by Kleinberg [1], PageRank algorithm [2], Vector Space Model, Latent Semantic Indexing and others.

The paper presents an adapted HITS algorithm and its implementation as a program for synonyms and related terms search in a corpus with hyperlinks and categories (Wikipedia). Since the developed algorithm uses a link structure (hyperlinks in texts), it is language independent.

In order to select text resources and software libraries the following criteria were taken into account.

- It should be open, or at least free.
- The developed program and used libraries should be operating system independent.
- Some library for visualization of synonym search result is needed.

The Wikipedia (as the text corpus) and TouchGraph WikiBrowser[1] (visualization library) were selected. The proposed synonym search algorithm and the developed program[2] could be used in order to extend query in a search engine, or as an assistant for forming a dictionary of synonyms.

### 1.1. Wiki resource: Wikipedia

Wiki is a type of sites that provide a simple way to add and modify pages for users. It is specially designed for collaborative work[3]. Wiki is also a software package at server side enabling one to add/modify Internet pages content via Internet browsers. Wiki language supports hyperlinks (to create links between wiki pages), it is more human readable than HTML and more safe (there are no JavaScript and Cascading Style Sheets).

The wiki resource Wikipedia is used in this work as a text corpus. Wikipedia[4] is a free online encyclopaedia in English, Russian, etc.

The research and discussion of Wikipedia resource are presented in [3]. A Wikipedia article[5] is defined as a page that has encyclopedic or almanac-like information ("almanac-like" being lists, timelines, tables or charts). Wikipedia text corpus has the following features valuable for synonym search:

- The texts classification is defined via categories [4]. Article authors select and assign most suitable categories to the article. It is possible to create a new category and link it to other relevant categories.
- The encyclopedia contains a lot of articles related to different topics, related to modern topics, since the encyclopedia is updated every day.

### 1.2. Corpus requirements

The set of documents constitutes a corpus of the documents. Hyperlinks and categories are used to search for similar texts in the corpus. The texts in the corpus are characterized as follows:

1. Text documents include a set of keywords, if keywords are not given, the document title is considered as a set of keywords.
2. The documents refer to each other via hyperlinks. Every document has a set of out-links (hyperlinks to documents this document refers to) and in-links (hyperlinks this document is referred by). The hyperlinks are assigned by experts.
3. Every document belongs to a set of categories/topics. Belonging of a document to a category is defined by experts. Categories make up a tree-like structure, thus every category has a parent category (except the root) and child categories (except the leaves).

Wikipedia satisfies corpus requirements. The search of synonyms within such corpus could be presented in a strict mathematical formulation.

[1] http://www.touchgraph.com

[2] The open source program Synarcher is written in Java. It is available at (http://sourceforge.net/projects/synarcher)

[3] http://en.wikipedia.org/wiki/Wiki

[4] http://wikipedia.org

[5] http://en.wikipedia.org/wiki/Wikipedia:What_is_an_article

## 2. HITS algorithm

HITS algorithm searches Web pages that are relevant to a given query. One of alternatives to the HITS algorithm is the Page rank algorithm [2] incorporated in Google.com. Google description of PageRank is suitable for HITS[6]:

> PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important".

The notion of authority is used in HITS. Authority has meaning only in the context of a particular query topic. Link structure is used in HITS to identify page authority, since hyperlinks encode a considerable amount of latent human judgment. "*Authoritative pages* relevant to the initial query should not only have large in-degree (hyperlinks this document is referred by); since they are all authorities on a common topic, there should also be considerable overlap in the sets of pages that point to them. Thus, in addition to highly authoritative pages, it is needed to find what could be called *hub pages*: these are pages that have links to multiple relevant authoritative pages" [1].

It is proposed to use link structure in HITS in order to satisfy similar-page queries. Using the notion of hubs and authorities, it is possible to solve the issue of page similarity. Given a page $s$ that is of interest, the authorities in the local region of the link structure near page $s$ can potentially serve as a broad-topic summary of the pages related to $s$.

The idea of link structure usage is adapted for synonym search in Wikipedia.

## 3. Adapted HITS algorithm

HITS algorithm was adapted in order to use the additional characteristics of documents (pages) provided by the corpus.

Since pages include a set of keywords[7] (page title in Wikipedia case), the keywords of similar pages could be considered as synonyms for the keywords of the source page. So the synonym search problem can be presented as a problem of similar pages search. In turn, the issue of a similarity (between pages with hyperlinks) can be formulated as a problem of *a search of similar graph vertices* based on hub and authority notion.

### 3.1. Problem statement

Given directed graph $G = (V, E)$, the vertex set $V$ (documents), the arc set $E$ (links), the source vertex $s$. For each document $v$ there are two lists: $\Gamma^+(v)$ includes documents which are referred by the source document $s$, the list $\Gamma^-(v)$ includes pages referring to the source document. For each vertex there are two weights, authority and hub: $\{v \in V : a_v, h_v \in \mathbb{R}\}$.

It is needed to find the set $A$ of vertices that are (i) *authority vertices* to the source document $s$ (i.e. value (1) for A is higher than for other subsets of vertices of the same size $N$), (ii) *similar vertices* to the source $s$ (i.e. there exists a vertices set $H$ such that for each vertex from $A$ there are vertices in $H$ that simultaneously have out-links to $s$ and to vertices from $A$ (2)), (iii) the set $H$ consists of hub vertices (i.e. value (3) for H is higher than for other subsets of vertices of the same size $M$). The aim is to select the set $A$ of authority vertices and the set $H$ of hub pages corresponded to (4).

$$\sum_{v \in A} a_v \xrightarrow[A \subset V, |A|=N]{} \max \quad (1)$$

$$A \subset V, H \subset V, \forall a \in A \quad \exists h \in H : \Gamma^+(h) \ni \{s, a\} \quad (2)$$

$$\sum_{v \in H} h_v \xrightarrow[H \subset V, |H|=M]{} \max \quad (3)$$

$$k \cdot \sum_{v \in A} a_v + (1-k) \cdot \sum_{v \in H} h_v \xrightarrow[A \subset V, H \subset V]{} \max, k \in [0, 1] \quad (4)$$

### 3.2. Algorithm

The algorithm input parameters are $s \in V$; $t, d, N, C_{max} \in \mathbb{N}$; $\varepsilon \in \mathbb{R}$, where

- $s$ – the source document for which similar documents are sought for.
- $t$ – root set volume (number of documents included in the root set of documents);
- $d$ – the parameter defining base set[8] volume ($d$ in-links for each document from the root set will be added to the base set, see more details in [1]);
- $N$ – number of similar words to be found (the same: number of similar documents to be found);
- $C_{max}$ – maximum weight of the cluster (number of documents in the cluster, number of categories, etc. are taken into account), where cluster is a set of documents which have common hyperlinks to other documents and *to categories*;
- $\varepsilon$ – the iteration error.

Steps of the adapted algorithm:

1. The source document $s$ refers to a set of documents that (along with $s$) form the root set of documents ($\leq t$ documents should be included).
2. For each document in the root set of documents: all out-link documents and not more than $d$ in-link documents are included in the base set (it is the same step as in HITS algorithm).
3. Document weights are calculated iteratively in accordance with HITS formulas. The iterations stop when the iteration error[9] is less or equal to $\varepsilon$.

   HITS uses two values for rating pages: the authority value $a_j$ and the hub value $h_j$, which are defined in terms of one another in a mutual recursion ($E$ is the set of hyperlinks):

   $$a_j = \sum_{i:(i,j) \in E} h_i \quad (5)$$

   $$h_j = \sum_{i:(j,i) \in E} a_i \quad (6)$$

   An authority value (5) is computed as the sum of the hub values that point to that page. A hub value (6) is the sum of the authority values of the pages it points to.

[6] http://www.google.com/technology

[7] See the $1^{st}$ characteristic of corpus documents in 1.2.

[8] Base and root set forming is described below.

[9] Sum of changes of hub and authority weights of all vertices after one iteration.

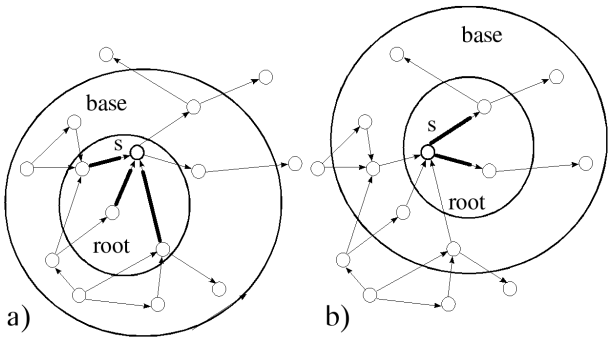


Figure 1: Root and base set forming in (a) HITS and (b) adapted HITS algorithm

4. Hierarchical clustering algorithm is applied to base set in order to cluster documents into group corresponding to different topics (this step is absent in Kleinberg algorithm). The clustering algorithm takes into account document weights (calculated on previous step) and hyperlinks structure of documents and categories.

5. For each cluster of the base set: select set $A$ (contains $N$ documents) such that (i) set $A$ consists of authority documents (i.e. the total document authority weight is large enough compared to other subsets of documents of the same volume) and (ii) for each document $a$ in $A$ there are hub documents which refer to $a$ and to the source document $s$[10].

HITS and the adapted algorithms are different in the way of root set forming (Figure 1). In HITS a root set consisting of $t$ pages that *point to* $s$ (a) is assembled using relevant pages found by a search server (Altavista). In the adapted algorithm, pages, which are ponted by $s$, form the root set (b); then the root set grows into a base set as in HITS; and the result is a subgraph $G$ in which hubs and authorities are identified[11].

The adapted algorithm also introduces the ideas of clustering base set of pages (into documents groups corresponding to different topics) (i) using calculated hub and authority weights and (ii) taking into consideration topics of the documents[12].

## 4. Implementation

Adapted HITS algorithm was implemented in the program titled Synarcher.

Data for the program Synarcher are Wikipedia in MySQL format (the structure of tables corresponds to the requirements of the MediaWiki[13] system). Though Synarcher was tested with English and Russian Wikipedia data only, it is supposed that Synarcher can perform search within other wiki resources, which are based on MediaWiki, as well[14].

The Synarcher was tested on Windows XP and Mandrake Linux. Running the Synarcher on client side requires a Java Runtime Environment (ver. 1.3.0 or above). The server should provide an access to Wikipedia resources via MySQL, Apache, and MediaWiki software. The currently public available Wikipedia servers (e.g. http://en.wikipedia.org) could not be used directly since the current version of Synarcher requires intensive computations. Hence a locally installed version of Wikipedia was used.

The results of synonyms search are presented in the form of a table and graph (Figure 2). The graph presentation of synonym search results is based on TouchGraph WikiBrowser V1.02. The user is presented with a split view, with a conventional html browser on the left side of the screen, and a graph of a local region of the wiki on the right.

The link graph consists of nodes whose labels correspond to hypertext pages (labels are the names of articles in Wikipedia), connected by directed edges corresponding to hyperlinks between the pages. User can expand node, hide node's neighbours, or rate node as a synonym for the requested word.

The user enters a word, and the program performs automatic search of its synonyms. Then, the user explores the result graph (base set of articles for the source article $s$) and rates (option rate/unrate) words, which are synonyms from the user's point of view (this is an interactive part of the work with graph or table). Search parameters[15] and synonyms selected by the user are stored at client side.

## 5. Experiments

The local versions of English and Russian Wikipedia (corresponding to the online version on March 8, 2005) were used in experiments. The English encyclopedia contains 901,861 pages, and 18,380,035 links, Russian encyclopedia contains 30,161 pages, and 468,771 links.

It should be noted, that the search for synonyms is not fully automated, since the program forms the list of related words, but many of words are not synonyms to the source word. The additional interactive search (in a graph or in a table) is needed. Thus, the program provides filtering of related words (and potential synonyms), which serve as raw data for an expert.

The expert has identified (during an interactive search in Synarcher) seven synonyms for the word *robot*: *android,* **golem**, *homunculus, domotics, replicant, sentience, parahumans*[16] (synonyms presented in the thesauri WordNet[17] or Moby[18] are written here and below in boldface). WordNet 2.0 contains only two synonyms for the word *robot*. They are **automaton** and **golem**.

With the help of Synarcher four synonyms for the word *astronaut* were identified: **cosmonaut**, *taikonaut, spationaut*, **space tourist**. WordNet proposes two synonyms: **spaceman, consmonaut**. There is no entry *robot* in Moby, but there are 79 related words for the word *astronaut* in it, six of which can be considered as synonyms: **aeronaut, cosmonaut, pilot, rocket**

[10] See formulas (1)–(4) in problem statement, sec. 3.1.

[11] The advantage of the approach presented in (Figure 1a) is that the root set includes hubs, so the base set can include authorities. The disadvantage is that a lot of pages can refer to the source page $s$, root set volume should be constrained by $t$, and some hubs can be missed.

In (Figure 1b) number of out-links (the volume of root set) is limited, since is constrained by size of the document $s$. But it is not evident that the base set will include authorities. Experiments for comparison of the approaches are to be done (measures to compare results should be developed).

[12] See the $3^{rd}$ characteristic of corpus documents in 1.2.

[13] MediaWiki (`http://www.mediawiki.org`) is a free software package originally written for Wikipedia.

[14] See sites using MediaWiki (`http://meta.wikimedia.org/wiki/Sites_using_MediaWiki`).

[15] See parameters of algorithm in sec. 3.2.

[16] Explanation of the meaning of words can be found at (`http://en.wikipedia.org`).

[17] http://wordnet.princeton.edu

[18] Moby Thesaurus List by Grady Ward (`http://www.dcs.shef.ac.uk/research/ilash/Moby`).

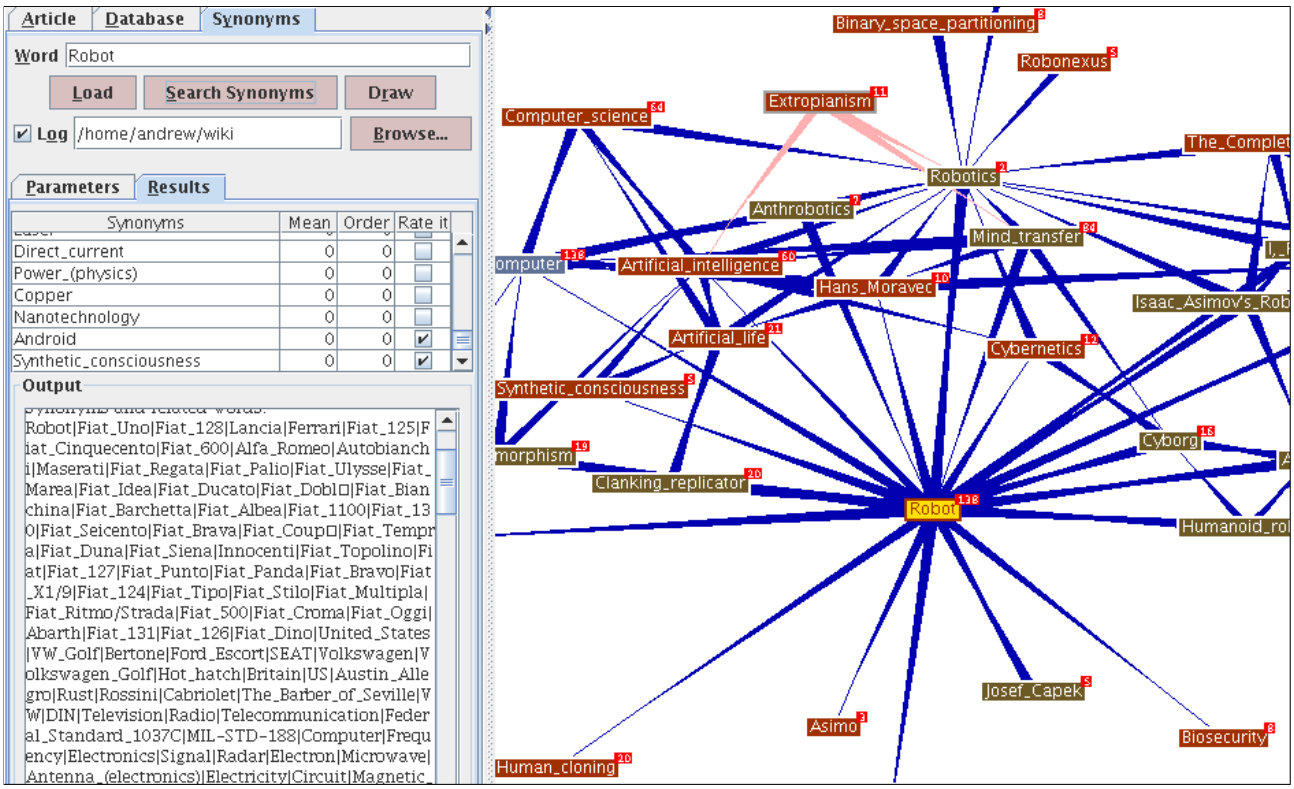


Figure 2: The found synonyms and related terms (for the word Robot) presented in the table and graph (English Wikipedia)

**man, rocketeer, spaceman**. Thus, two and six synonyms for the words *astronaut* and *robot* were found, which are absent in thesauri WordNet and Moby.

Small size of the current version of the Russian Wikipedia does not allow one to find as many synonyms (within related terms gathered by Synarcher) as the English Wikipedia.

## 6. Future work

Future work includes estimation of the quality of synonym search, e.g., by comparing the found set of synonyms against the WordNet and Moby thesauri.

The hierarchical algorithm is planned to be used to cluster the result set of words related to different problem areas. It is possible for wiki resources, since every article has one or more assigned category types (e.g. the article "QNX" has the following categories in Wikipedia: Unix, Computing platforms, Embedded operating systems).

## 7. Conclusions

Adapted HITS algorithm for search of synonyms and related terms in the corpus with hyperlinks and categories is presented. The algorithm is implemented in the program Synarcher. English and Russian Wikipedia are used as the text corpus.

The experiments show that successfull search of new synonyms is possible with Synarcher. New synonyms, which are absent in thesauri WordNet and Moby, for the words *robot* and *astronaut* were found. It could be explained by the properties of Wikipedia.[19]

The proposed algorithm could be applied to a query modification (expansion by synonyms) in a search engine. The program Synarcher can serve as an assistant in a synonym dictionary forming.

## 8. Acknowledgements

This work was partly supported by the Russian Science Support Foundation

## 9. References

[1] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999. [Online]. Available: http://www.cs.cornell.edu/home/kleinber

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," 1998. [Online]. Available: http://www-db.stanford.edu/~backrub/google.html

[3] T. Holloway, M. Bozicevic, and K. Borner, "Analyzing and visualizing the semantic coverage of wikipedia and its authors," 2005. [Online]. Available: http://arxiv.org/abs/cs/0512085

[4] J. Voss, "Collaborative thesaurus tagging the wikipedia way," 2006, collaborative Web Tagging Workshop. [Online]. Available: http://arxiv.org/abs/cs/0604036

[19] The English encyclopedia contains a lot of articles (998,470 in English on March 1, 2006) related to different topics (science, art, politics, etc.). It contains articles related to the modern topics, since the encyclopedia is updated every day.