VISUALIZATION

Alexander Babaev

Department of Programming Technology,
Faculty of Applied Mathematics and Control Processes,
St. Petersburg State University,
Universitetsky pr. 35,
St. Petersburg, Russia
koptilo@mail.ru

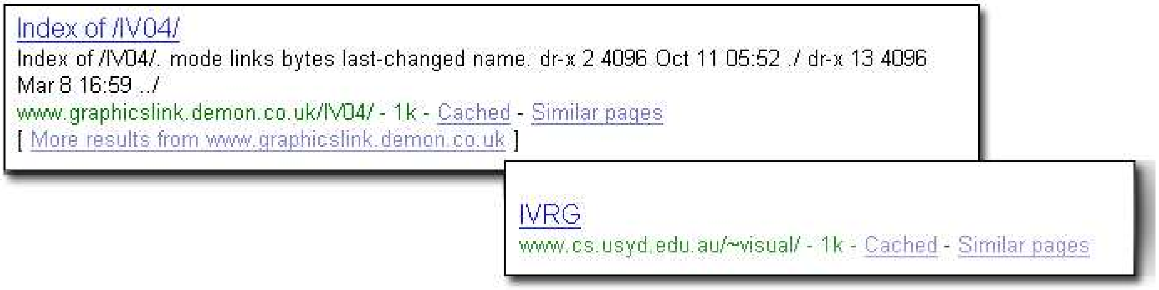
Abstract

This paper examines some aspects of information needs and human information perception in terms of modern search engines. There is also a short overview and comparison of popular nowadays internet search systems.

To avoid some misunderstandings, the author should notice that this paper is partly a compilation of the material that referenced in the last section.

I. Introduction.

When we speak about information search and especially internet search the first thing that comes into mind is **google**-like search systems**.** These Boolean-algorithm based engines give in most cases acceptable results; however user should waste some time to find it among all returned documents. Sometimes, if you don't know exact name of document, defining a couple describing words in a query could result in appearing hundreds of barely connected matches. As description of match such engines use a part of found document with highlighted words from query, page title and link to the article. Figure 1 shows some results for the query "information visualization" returned by Google in first 20 matches.



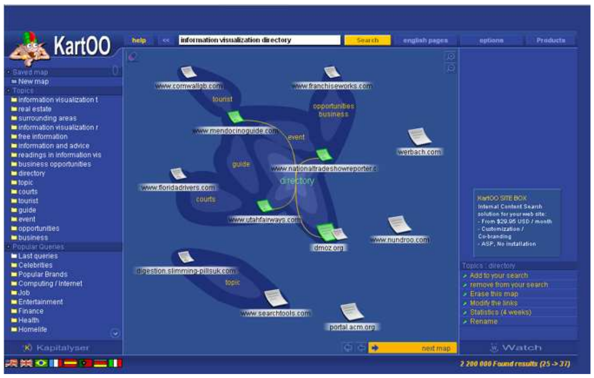
Index of /IV04/
Index of /IV04/. mode links bytes last-changed name. dr-x 2 4096 Oct 11 05:52 ./ dr-x 13 4096 Mar 8 16:59 ../
www.graphicslink.demon.co.uk/IV04/ - 1k - Cached - Similar pages
[More results from www.graphicslink.demon.co.uk]

IVRG
www.cs.usyd.edu.au/~visual/ - 1k - Cached - Similar pages

Figure 1. Some search results for query "information visualization"

As in this case, finding information in popular modern search engines often becomes surf between badly described matches.

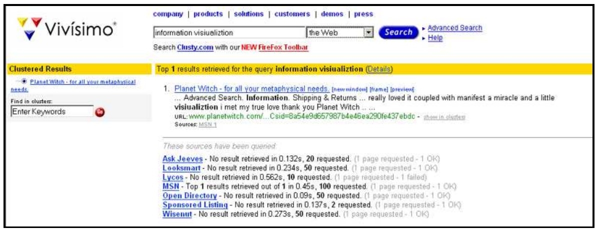
However there are some approaches to make search more effective and to provide user with friendlier interface than text-based ranked list. These approaches are often based on visualization and presentation of relationships between documents, terms or user query. In some approaches (like vivisimo.com) information clustering is used so that user may choose the segment of search space without the need to browse through all hits, in others (e.g. kartoo.com) modern technologies like *flash* are used to represent traditional 10 hits as fancy 3D globes, but without summaries. However, in both cases Information retrieval process and especially visualization (in meaning of usability and assistance to user) are not good enough to compete with google`s simplicity and power. For example, using query "information visualization" in Vivisimo.com results in one hit quite distant to expected results (need to mention that Vivisimo only uses results from other engines like MSN, thus being only a visualization shell), some better results were obtained from kartoo.com, but despite all its beauty, kartoo is still far away from "intuitive" interface and engine itself is less powerful, comparing to google.



information visualization directory

2 200 000 Found results (25 -> 37)

Figure2. Kartoo



information visiualiztion | the Web

Search Clusty.com with our NEW FireFox Toolbar

Top **1** results retrieved for the query **information visiualiztion** (Details)

Planet Witch - for all your metaphysical needs.

1. Planet Witch - for all your metaphysical needs. [new window] [frame] [preview]
... Advanced Search. **Information**. Shipping & Returns ... really loved it coupled with manifest a miracle and a little **visiualiztion** i met my true love thank you Planet Witch ... ...
URL: www.planetwitch.com/...Csid=8a54e9d657987b4e46ea290fe437ebdc - show in clusters
Sources: MSN 1

These sources have been queried

Ask Jeeves - No result retrieved in 0.132s, **20** requested. (1 page requested - 1 OK)
Looksmart - No result retrieved in 0.234s, **50** requested. (1 page requested - 1 OK)
Lycos - No result retrieved in 0.562s, **10** requested. (1 page requested - 1 failed)
MSN - Top **1** results retrieved out of **1** in 0.45s, **100** requested. (1 page requested - 1 OK)
Open Directory - No result retrieved in 0.09s, **50** requested. (1 page requested - 1 OK)
Sponsored Listing - No result retrieved in 0.137s, **2** requested. (1 page requested - 1 OK)
Wisenut - No result retrieved in 0.273s, **50** requested. (1 page requested - 1 OK)

Figure 3. Vivisimo

In many respects visualization of information nowadays is back in the Stone Age [1]. So what aspects make visualization successful?

II. Information needs.

Successful search in many aspects depends on a type of information you are searching. For example searching news, "blogs" entries, multimedia or maps in basic search engine in most cases becomes a "search in search" when you need to understand whether and what given matches correspond your need or they have just a mentioning of a query words. For some needs google-like systems suggest solutions, so Google itself has options to search for specific document types like pdf, ppt, xls and other formats; Lycos, Yahoo and almost all big engines support image, audio and video search; some engines give ability to choose a "find in…" option – kartoo make search in different language zones, as does google, vivisimo suggest more interesting grouping – search in Web, News, Forums by using different engines as it was said before; there are also options to make search in blogs. However, all this approaches are split and no system can say that it "feeds" all informational needs. But what are they? To solve a problem we must first decide what kinds of information needs are there now.

And in this respect visualization and Information retrieval itself are quite young, because there are not many researches in this sphere at this moment. But processing existing researches we may categorize information needs and ways to “feed” this needs in the following groups:

Information need	Solution
Known-item	Search system, site index
Exploratory/orientation	TOC/site map, guide, top levels of hierarchy
Open-ended	Guide, hierarchy, search wizard, easy switching between search and browse, collaborative filtering
Selective research	Search system, filtered results through use of search zones
Comprehensive research	Search system, expanded results through use of thesaurus

Table 1. Information needs classification [3]

As we may see, suggested solutions don’t have common items, thus search in each case requires specific way or specific engine. However, there are approaches to create universal visualization engine independent to search system. One kind of such approaches are numerous shells like already mentioned Vivisimo, Ez2find, dogpile, which use their interface to process information provided by other engines. Another approach is creation of some intermediate language, which will allow interaction between any search and any visualization engine. Such language called IVL (Information Visualization Language) is based on xml and its common working schema is quite simple and looks like this: “when the user issues a query (1), a query parser module translates the query to the selected search engine syntax (2). The results from the query can be in IVL format or not depending if the searcher supports IVL or not. If the searcher does not support IVL, a transformer will map the internal format of the searcher to IVL (3). Finally if the interface does not support IVL, a transformer will perform another transformation to generate the visualization’s format (4)” [4].

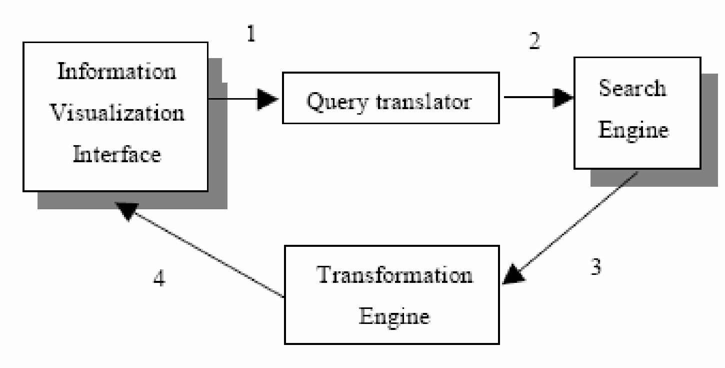


Figure 4. IVL Software Architecture

Combining most popular interfaces with most popular search engines can give user more flexible search. However in all these cases a problem of choosing engine (and so – defining information need) still stays a user’s problem.

Until this moment only *internet* information search engines were subject of our research, however, when speaking about *enterprise* engines in terms of information needs, most of named problems are less actual. For example, specialist working with medical articles database will probably use it to find some appropriate medical information and will not use it to find his friend’s home address or space star map. Thus, knowing the sphere of provided documents, a better visualization can be given. However, as enterprise search has fewer needs common with global internet search, it has its own specific problems connected with greater requirements to search “precision”, which are not discussed in this article.

III. Information Perception.

Now let’s see what is important to build good visualization in aspect of human information perception. First of all we should mention Miller`s magical number seven. In 1956 George Miller published his article “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information” where he estimates capacity of human short-term memory to 7 objects [5]. In other words, having long list with common objects (like query search results - ranked list) we may simply “sink” in numerous matches. To prevent it, “to minimize cognitive load and maximize the information assimilated subconsciously”, interface should use available visualization dimensions more effective. For instance, speaking about “display dimensions” the following heuristics should be mentioned [1]:

· **Attribute Resolution**: For representation involving a single output dimension only six or seven distinctions can be handled without conscious processing.

· **Number of Attributes**: It seems that it is pointless to visualize more than six or seven features or attributes to distinguish ‘data facts’, and even then the resolution that can be subconsciously processed and recalled may be limited to only two or three distinctions per feature.

· **Explicit and Implicit Grouping**: It is useful to represent data facts in such a way as to allow the user to subconsciously group and recode. Whilst clustering techniques can be used to explicitly recode and limit the amount of detail, visualizations showing natural clusters can convey the same information implicitly, as long as appropriate dimensions are displayed.

· **Natural Interactivity**: A user should be able to interact with the display in a way that leads to intuitively reasonable modification to the display (e.g. new views showing different perspectives or levels of detail).

· **Views and Cues**: When changing views we should provide cues to help the user ‘clear out’ the old information in the dimensions that are being reused. In addition it is helpful to have cues to clarify the relationships and continuity between views. Various animation techniques can serve one or both of these purposes. A common approach is to retain all data as context but have some in a higher resolution focus.

· **Sequential and Parallel Presentation**: Distinctions that may not be salient in a simultaneous presentation may become salient when it is animated, so that time becomes an additional dimension available to contrast data objects or present or reinforce a specific attribute.

So now, having some requirements defined we should also mention evaluating methods and measures to define IR visualization interface quality. And first two criteria to mention here of course are standard IR measures:

1. **Recall**: a measure of how well the relevant results are represented in the data returned, being the ratio of the number of relevant retrievals to the total number of relevant documents in the collection.

2. **Precision**: a measure of how much of what is returned is relevant, being the ratio of the number of relevant retrievals to the total number of retrieved documents.

Unfortunately recall and precision suffer from a number of limitations, and assume the return of a single unranked set of results for a single query. When used with ranking the problem thus becomes when to cut off the returns, so there is a tradeoff between returning more results with the hope of increasing recall and fewer results with the hope of increasing precision.

When clustering is carried out there is a more complicated problem of assessing the utility of clusters, and here multiple manually developed classes may be compared with automatically determined clusters. Furthermore, a visualization interface involves providing multiple viewpoints and allowing users to cull the results interactively. Clearly other factors must be taken into account before we can sensibly apply and interpret recall and precision. Indeed, there are many problems with these as accuracy measures and so some additional evaluating properties should be used. For that purposed the following measures were developed:

3. **Bookmaker accuracy**: to what extent are the results due to correct or incorrect use of information rather than random guessing?

4. **Time**: given a single retrieval task what time is taken, including system and user time, to achieve that task?

5. **Number of interface interactions**: given a single retrieval task how many times does the user interact (e.g. click, drag, etc…) with the graphical interface?

6. **Number of refinements**: how many times has the query been refined?

7. **User opinion**: the opinion of users should be solicited under controlled conditions to help capture factors such as the intuitiveness and friendliness of the interface.

8. **Cognitive load**: a user's mental load in using the interface to achieve a specific search/retrieval task.

Cognitive load is directly influenced by the design of the user interface and will typically be measured by assessing how effectively a user can use the interface concurrently with other tasks or distractions. Here we are not necessarily seeking to minimize cognitive load. Rather the ideal is to maintain a level of cognitive load such that the task is not trivial or boring, yet does not overload the user to the point where the tool is difficult to use or error rates increase.

9. **Number of errors**: how many mistakes did the user make when using the interface?

10. **Learning curve**: how quickly can a novice user learn to become proficient in using the interface?

11. **Effective use of screen real estate**: does the visualization effectively use the maximum amount of screen area available?

12. **Number of results displayable**: how many results can be effectively displayed to the user in a given area of the screen?

13. **Mode of use**: what task is the interface being used for? e.g. Searching for answers to a specific question, a specific document or a specific reference.

14. **Multi-session support**: does the interface support use of usage history or feedback usable adaptively in other searches by the same or another user?

15. **Significance**: Have we enough data/subjects/trials for our results to be statistically significant and what is the probability that our results are due to chance?

16. **Bandwidth**: what is the trade-off between server load, client load, and network load?

So, using this classification, comparison of different search engines visualization effectiveness is possible and, following balance in these classifications, building even more effective visualizations becomes more target-oriented.

IV. Current situation.

So now, having defined requirements for good visualization system, let's look on popular engines closer, through the scope of mentioned above criteria. First of all, basic model of information access process [6] nowadays looks as shown on fig.5:

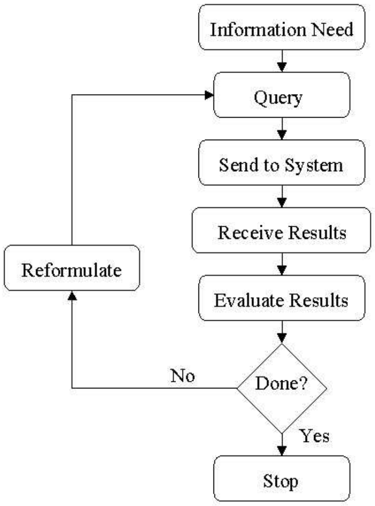


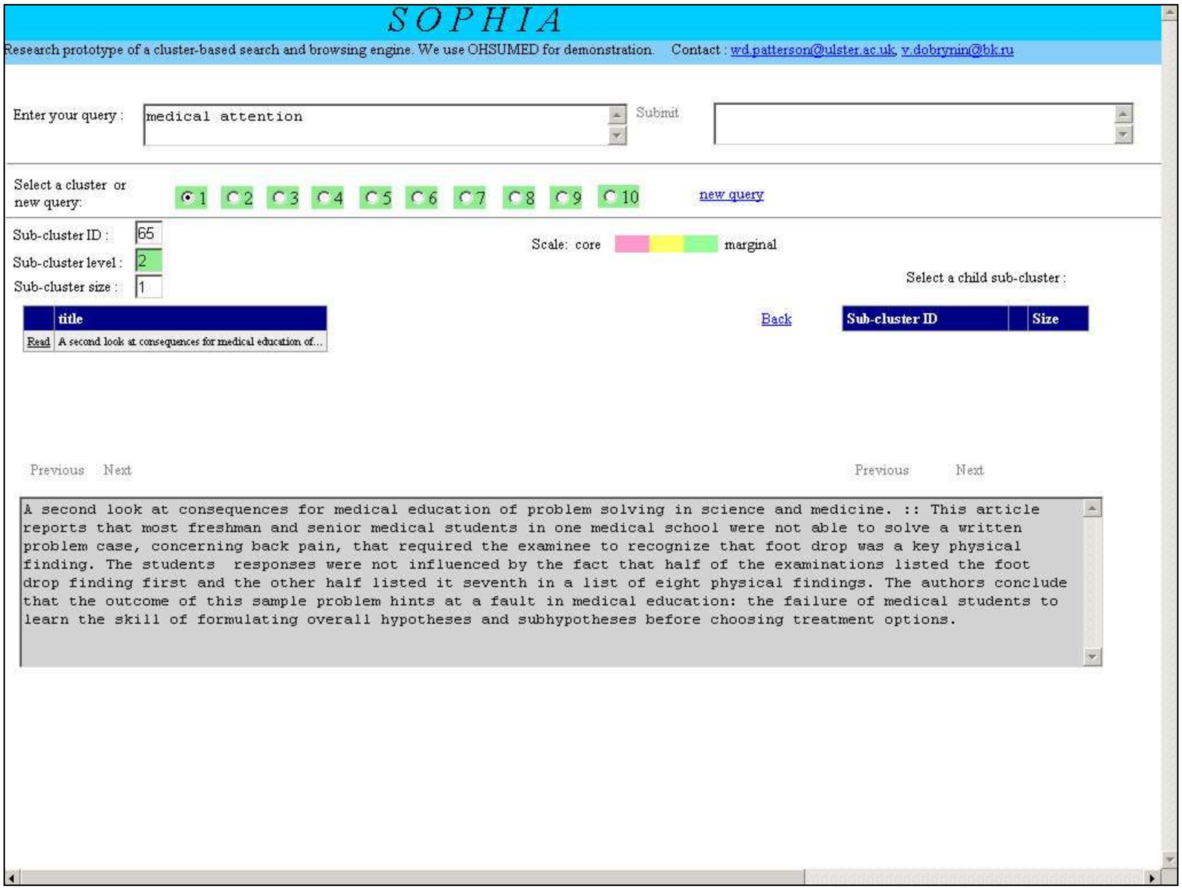
Figure 5. Diagram of the standard model of information access process

So, unsuccessful search leads to reformulation of query, thus repeating almost all steps in this chain. This situation is often met in google-like engines, when without any ability to walk through found documents topics the only way to change the direction of search becomes query changing. In other respects these engines have the following advantages: most of them support a kind of *Selective research* – images/multimedia/news/answers/catalogs search, however these powerful functions are often hidden behind the “advanced” link. If to speak of such systems in terms of information perception, first we should mention ineffective document space usage – so to show 10 matches google generates html page where even at very high screen resolutions scrollbar stays an object of great necessity, while much space is devoted to “sponsored links” or simply free white space, also no advanced interactivity or any kind of results grouping are available – all hits are static ranked-list. However, the simplicity of those engines results in good *learning curve*, low *number of errors* while working with interface and (in case of experienced users) low *time* and *cognitive load* to reach search target. Also, being a thin-client in this case results in low *client-load.*

Kartoo.com. This search engine differs from all other greatly, most of all by its visual part. Traditional 10 matches are represented here graphically, using flash technology, with short google-like description appearing in special window as user moves a mouse over a match icon. In this interface a much better space “administration” can be seen. However, becoming more effective, it lost simplicity and “intuitiveness”, so to become familiar with it you have to spend some time over documentation to understand all its features. Using a flash-based client probably lowers network traffic, as there is no need to send data, embedded in html text. One more kartoo`s advantage is that it uses some basic grouping / clustering approach. Matches connected by common words (these connections are also shown in the working area) are put in one cluster, “entering” such cluster makes automatic querying of type: old query + selected cluster name.

Vivisimo.com. This visualization engine is focused especially on clustering search results. It has a simple explorer-like interface – in the left panel created clusters are shown and right panel shows matches inside selected one. Thus, we see that interface is “intuitive” and requires no special training. By viewing suggested clusters user makes fewer clicks to the target, as there is no need to browse wide range of results and all matches not corresponding to required information need may simply be ignored. Such approach is not new in IR concepts; however, it is very effective.

The practice of using clusterising and cluster search is quite rich in enterprise engines, it has been used for years and, of course, there is much advancement in this sphere. So as we are interested in visualization, let’s look at powerful and perspective enterprise search engine SOPHIA through the scope of our research.


SOPHIA

Research prototype of a cluster-based search and browsing engine. We use OHSUMED for demonstration. Contact : wd.patterson@ulster.ac.uk, v.dobrynin@bk.ru

Enter your query : medical attention

Select a cluster or new query: ● 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Sub-cluster ID : 65

Sub-cluster level : 2

Sub-cluster size : 1

Scale: core 🟥🟨🟩 marginal

Select a child sub-cluster :

	title
Read	A second look at consequences for medical education of...

Sub-cluster ID	Size

A second look at consequences for medical education of problem solving in science and medicine. :: This article reports that most freshman and senior medical students in one medical school were not able to solve a written problem case, concerning back pain, that required the examinee to recognize that foot drop was a key physical finding. The students responses were not influenced by the fact that half of the examinations listed the foot drop finding first and the other half listed it seventh in a list of eight physical findings. The authors conclude that the outcome of this sample problem hints at a fault in medical education: the failure of medical students to learn the skill of formulating overall hypotheses and subhypotheses before choosing treatment options.

Figure 6. SOPHIA

One of a visualization problem in enterprise search is that it is committed in a space of documents of common topics, and thus it has to be more cognitive to searcher comparing to what we have in internet search. Another problem is that document base may contain great number of articles and some queries may result in hundreds or thousands matches, so a good representation of these results is also required. So how does SOPHIA solve these problems?

In this article we will not discuss SOPHIA`s algorithm itself, but it is to be noted that a given document base is split into clusters which are then searched for match using special non-Boolean-based algorithm which allows to find more exact hits. So let's look at SOPHIA visualization (which is at this moment in "alpha" state). Fist thing that comes into mind – there is no huge lists. Indeed, this system allows walking through clusters and found data easily – as search is complete user should choose one of 10-top clusters containing found articles, as it is done navigation inside cluster continues by browsing found articles or sub clusters. Through there is not much additional functionality in this interface at this time, all required tasks are committed easily. However, there is still much work to do with it.

V. Conclusion

This paper examines modern search engines in concept of visualization, in meaning of assistance to user to find required information depending on his *information needs.* Different aspects of human *information perception* are also discussed here. Finally, clustering algorithms are proposed to be a solution to provide better assistance to user in search. Also, some aspects of enterprise system SOPHIA visual interface are described and discussed.

References.

[1]. David M. W. Powers, Darius Pfitzner "The Magic Science of Visualization"

[2]. "Advice Engine"
http://www.noodletools.com/debbie/literacies/information/5locate/adviceengine.html

[3].http://www.louisrosenfeld.com/home/bloug_archive/000139.html

[4]. Omar Alonso, Ricardo Baeza-Yates "Integration of Visualization with Search Engines"

[5]. George Miller "The magical number seven, plus or minus two: some limits on our capacity for processing information."

[6]. Ricardo Baeza-Yates, Berthier Ribeiro-Neto "Modern Information Retrieval", chapter 10.

[7]. www.google.com

[8]. www.kartoo.com

[9]. www.vivisimo.com

[10]. Dobrynin V., Patterson D, Rooney N., Contextual Document Clustering, Lecture Notes in Computer Science № 2997, Advances in Information Retrieval, 2004 year, p.167-180.