Searching the Greek WWW

Andreas Veglis[1], Andreas Pomportsis[2]

[1]Media Informatics Lab, Department of Journalism & MMC
[2]Department of Informatics
Aristotle University, 54124 Thessaloniki, Greece
Email: veglis@jour.auth.gr

Abstract: The amount of information available on the WWW is vast and growing at a staggering rate. The result of this growth was that users were unable to explore the vast recourses of the Internet. The answer to that problem was the search engines. Internet search engines first appeared in the mid-90s, but have, in a few years, made themselves part of our everyday lives. Following the global trend greek WWW has grown considerably in the last eight years. This paper presents an overview of the existing search engines that index the greek web pages, and attempts an initial comparison. Our study includes global and greek search engines. The comparison involves search engines' characteristics, relative size, overlapping, unique results and dead link analysis.

1 Introduction

The potential commercialization and global diffusion of the Internet after 1992 has established the Internet as a dominant means of communications [1-2, 10]. In the recent years the number of Internet users across the world has increased with surprising speed. People can access information and communicate with others without being constrained by space and time. The most dominant Internet service is the World Wide Web (WWW) [2].

The WWW became a major force in computer-mediated communication in 1995. This Internet service was quickly adopted in every human activity. Its possibilities are enormous and among other include interactivity, multimedia features, possibility of regular updates, and accessibility to archives [3-4].

As a source of immediately accessible data on current events and developments, in an extensive variety of fields, there is nothing, certainly in quantitative terms, to rival information stored electronically on Internet servers. No library could possibly afford the acquisition of as many hard-copy equivalents of electronic newspapers, mailing lists, Usenet newsgroups, and electronic magazines, as are available over the Internet, nor could they afford the costs of ordering and archiving such a volume of materials [1].

The amount of information available on the WWW is vast and growing at a staggering rate. The result of this growth was that users were unable to explore the vast recourses of the Internet. The answer to that problem was the search engines. Internet search engines first appeared in the mid-90s, but have, in a few years, made themselves part of our everyday lives. Today it is hard to imagine using the Internet with-

out them. Of course we often hear complaints, but we must always have in mind that sometimes the information isn't out there at all, and so search engines simply cannot help us. The WWW does not contain the answers to everything [5-10].

Following the global trend greek WWW has grown considerably in the last eight years [7]. As a consequence some greek search engines have appeared. In addition some well known global search engines have started indexing the greek WWW. This paper presents an overview of the existing search engines that help users search greek web pages, and attempts an initial comparison.

The rest of the paper is organized as follows: Next we present some details about the use of the Internet in Greece. A discussion concerning global and greek search engines can be found in the following section. This is followed by a comparison of the search engines. Concluding remarks can be found in the final section.

2 The Use of Internet in Greece

Currently 655 million people worldwide have access to the Internet. Internet usage is seeing an annual rise of about 30%. The Internet is by far more common in the EU than in Greece. Five EU countries (Germany, UK, Italy, France, and Spain) are among the top 15 counties with the most Internet users. Today, it is a daily routine for many more EU citizens than Greeks to use the Internet, i.e. to send an e-mail, to surf the web, to book travel, to order goods or services, to look for information whether is text, graphics or video. The electronic commerce is expected to attract even more users. Surveys indicate that 1 out of 3 homes in Greece own a personal computer. Also 19,3% of the Greeks are accessing the Internet. The number of Internet users has increased by 91% in the last year (see table 1). Two out of three computer users have access to the Internet. The Internet usage is more popular between young people. More precisely 50% of young Greeks (15-24 years old) have access to the Internet. Experts estimate that by the year 2004, 50% of the population will be using the Internet [7].

2001		*2002*	
Use of PC	*Internet use*	*Use of PC*	*Internet use*
20%	10,1%	28,9%	19,3%

Table 1. PC and Internet usage in Greece.

Most of the greek users access the Internet many times per week. It is worth noting that 35% of the users is accessing the Internet on a daily basis. More details can be found in table 2 [7].

Every day	*Many times per week*	*1-2 times per week*	*Less often*
35%	27%	25%	13%

Table 2. Frequency of accessing the Internet by greek users.

The purpose for accessing the Internet varies. 27,7% of the greek users use the Internet for entertainment, 24,4% for work, 21,7% for electronic mail, and 14,9% for education (see table 3) [7].

Entertainment	*Work*	*e-mail*	*Education*
27,7%	24,4%	21,7%	14,9%

Table 3. How greek users use the Internet?

3 Search Engines

There are two fundamental methods for locating information on the web: browsing and searching [9]. Browsing is the process of following a hypertext path of links. Searching, on the other hand, relies on powerful software that seeks to match the keywords the user specifies, with the most relevant documents on the web. Effective searching, unlike browsing, requires learning how to use the search software as well as lots of practice to develop skills to achieve satisfactory results.

There are two different tools available for the users in order to locate information. One method, called web directory, was modeled on early Internet search tools, like Archie and Gopher [9]. The other method, called a search engine, drew on classic information retrieval techniques that had been widely used in closed proprietary databases but hardly at all in the open universe of the Internet [5-6].

Search engines are databases containing full-text indexes of web pages. When you use a search engine, you are actually searching this database of retrieved web pages, not the web itself. Search engine databases are finely tuned to provide rapid results, which is impossible if the engines were to attempt to search the billions of pages on the web in real time [9].

Search engines employ complex programs. They are consisted of three parts [5-6]:

Web Crawler or Spider: It is a program that finds and fetches web pages. The crawler finds web pages with two methods. Most search engines have an *add URL* form, which allows web authors to notify the search engine of a web page's address. The second method of web page discovery takes advantage of the hypertext links that exist in most web pages. When a crawler visits a web page it also visits all the links it includes.

The Index Engine: It indexes every word on every page and stores the resulting index of words in a huge database, typically in an inverted data structure. An inverted index is sorted alphabetically, with each index entry storing the word, a list of the documents in which the word appears, and in some cases the actual locations within the text where the word occurs.

The Query Processor: It is the most complex part of a search engine. It compares the search query to the index and recommends the best possible matching documents. It is consisted of several parts, including the primary query interface, the actual engine that evaluates a query and matches it with the most relevant documents in the search engine database of indexed web pages, and the results-output formatter.

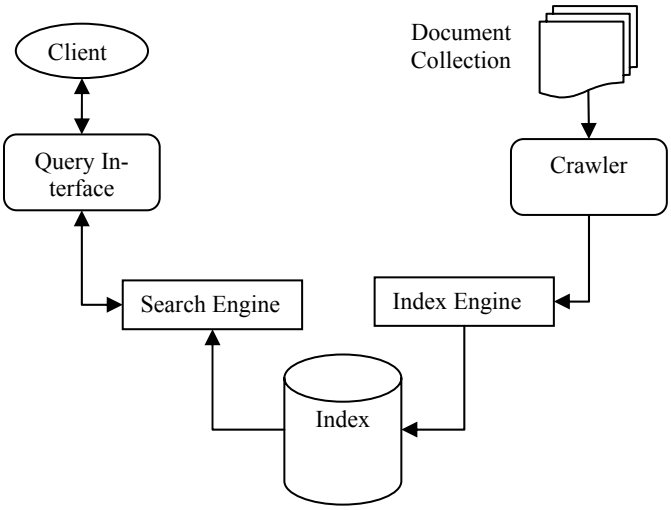

Fig 1. The operation of a search engine.

3.1 Global Search Engines

Today there are about 10 major search engines worldwide. Some of them are: Google, Lycos, Altavista, Teoma, MSN, AllTheWeb, and Inktomi. We have included in our study Google and Altavista because they are the search engines with the largest databases, and they index greek web pages.
Google (http://www.google.com): It claims to be the world's largest search engine. By accessing its index of more than 3 billion web pages, Google delivers relevant results to users all over the world, typically in less than half a second. Every day, Google responds to more than 200 million search queries. Google has developed an advanced search technology that involves a series of simultaneous calculations typically occurring in under half a second-without human intervention [12].
AltaVista (http://www.altavista.com): The search engine began its operation back in 1995. AltaVista added multilingual search with support for 25 languages in 1997. In 1999 it introduced multimedia (audio/video/image) search support. AltaVista was the first major search engine to introduce free Internet news search in 2001; and unveiled AltaVista Prisma, its powerful assisted search tool, in 2002 [13].

3.2 Greek Search Engines

Greek Internet includes many big and small search engines. Many of them offer similar characteristics as the foreign search engines. The most important greek search engines are: Trinity, POP, Phantis, and Anazitisis.

Phantis: It is the oldest and most famous greek search engine. It includes a rich thematic directory, and offers advanced search features. Although Phantis was created as a search engine, after a few years it became a portal [15].
Trinity: It is another good greek search engine. It is the search engine of the greek portal Pathfinder (http://www.pathfinder.gr). Trinity became operational in 1998. It includes a lot of advanced characteristics [14].
Anazitisis: It is the search engine of the ISP OTEnet (http://www.otenet.gr). It includes simple and advanced searching interfaces. Its results are satisfactory [17].
Pop: It claims to be the faster greek search engine. It includes an extensive database with more than 1,5 million web pages. Its interface is quite simple, but its results are very satisfactory [16].

Search Engines	*Boolean*	*Default*	*Proximity*	*Truncation*	*Fields*	*Limits*	*Sorting*
Google	-, OR	AND	Phrase	NO	intitle, inurl, more	Language, filetype, date, more	Relevance, site
AltaVista	+, -, AND, OR, AND NOT, ( )	AND, phrase	Phrase, NEAR	YES *	title, URL, link, more	Language	Relevance, site
Trinity	-	AND	NO	NO	NO	domain	Clustering by site
Phantis	+, -	AND	NO	YES *	NO	domain	Clustering by site
POP	+, -	OR	NO	NO	NO	NO	Some level clustering by site
ANAZITISIS	NO	AND	phrase	YES *	Title, url, description, keywords	domain	NO

Table 4. Characteristics of global and greek search engines.

3.3 Characteristics

In this section we perform a comparison of the characteristics [18-19] that global and greek search engines offer to their users. The results of our comparison are included in table 4. Based on the results we can make the following observations. Greek engines offer limited use of Boolean logic. Only one of them includes the use of a proximity operator and only half of them offer truncation capabilities. Only one greek search engine allows the user to specify the field were the search engine must search for the keywords. Finally the majority of the greek search engines offer domain limitation in searching and include result clustering by site.

From the above we can conclude that the characteristics of the greek search engines are somewhat fall short in comparison with the global search engines, especially with Google. But in general if we take into account the sizes of the global and the greek search engines, the findings are quite satisfactory.

4 Comparison

Measuring the constantly changing size of a search engine's database is a complex task. The size of a database is a very important issue in choosing the right search engine. Even with the best relevance ranking technology, search features, and user interface, a search engine cannot find the web page that does not exist in its database. With the continuously changing Web that offers new and changed information content daily, large databases become crucial tools for finding answers to questions beyond the very general and popular content offered by portals [18-19].

While a small, selective database may be more useful for extremely popular queries and very general topics, the strength of a large search engine database is that it can find web pages on less popular subjects, unusual products, distinctive keywords, smaller companies, small towns, and many other types of questions.

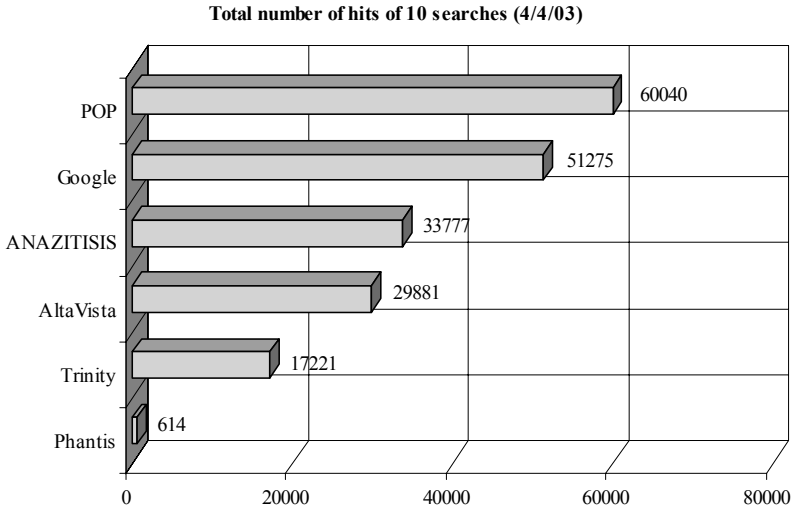

Fig 2. Relative size comparison.

The size comparison we have performed includes six search engines, with Google and AltaVista representing the global search engines. For the purpose of our analysis we have conducted 10 small single word queries. The 10 queries included words from current news and other popular subjects. Another approach would be to consider tenths or hundreds of carefully selected words (names, topics, etc.).

POP found more total hits than any other search engine (see fig. 2). But we must mention that Google produced more results in 5 out of 10 searches, and POP returned

more results in 4 out of 10 searches. AltaVista was placed fourth in the total number of results. As far as greek engines is concerned POP and ANAZITISIS appear to have the largest databases with Trinity coming third.

Next we compared the results of three small searches run on the same search engines. The three searches produced 82 results. Of those 82 web pages, 54 were found by only one of the six search engines while another 9 were found by only two. That means that more than half of all pages found were only found by one of the search engines, and not always the same one. Over 86% were found by three search engines at most. Each pie slice in the chart (fig. 3) represents the number of hits found by the given number of search engines. Based on the above we can conclude that there is little database overlap. That means that the database size of most of the search engines is small.

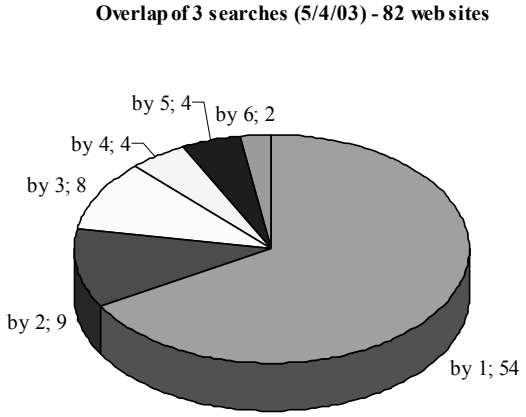

Fig 3. Overlap comparison.

For more detailed analysis, we plot in fig. 4 the percent of unique hits for each search engine. Unique hits were those URLs that were found by only one of the six search engines. The percentage figure refers to the percentage of unique hits as compared to the total hits that all the search engines found. POP found more than half (approximately 55%) of the unique results and Google only 17%.

Based on the results from the previous three searches, we include in table 5 the percentages of dead links among those 54 results. The dead links percentage column includes the 404 file not found error messages, the 401 access denied messages, the 403 forbidden errors messages, and various connection errors. The results show that Google produced the smallest percentage of dead links. 25% of the results from Altavista were inaccessible although Altavista returned a small number of results. The percentage of the dead links of the greek search engines ranged from 11% up to 26%. 17% of the results of POP, the greek search engine with the largest relative size, were inaccessible. In general we can conclude that greek search engines produced a considerable amount of dead links, that may be caused by the lower frequency of recrawling in comparison with Google.

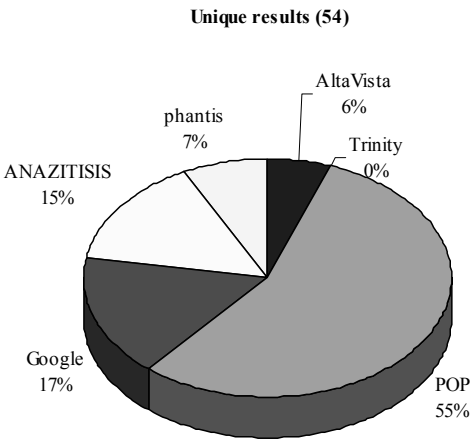

Fig 4. Unique results comparison.

Search Engines	*Dead Links %*
Google	7%
AltaVista	25%
Trinity	11%
Phantis	18%
POP	17%
ANAZITISIS	26%

Table 5. Dead link percentages.

5 Discussion

This study is a critical first step in understanding the present situation as far as searching greek web pages is concerned. The results indicate that greek search engines offer similar, to some degree, characteristics with the global search engines. The unexpected result came with the relative size estimation where POP came first and Anazitisis third (with Google the expected winner in second place.) This result was confirmed with the unique results analysis where POP produced 55% of the unique results and Google only 17%. Of course the percentage of dead links in POP's results is more than 2 times grater than Google's. Based on the above we may say that POP appears to index greek web pages better than Google. Of course POP has very few characteristics that limits a search (see table 4), thus making difficult for a user to browse through its results.

It is hard to come to any valid conclusions when many contradicting and subjective criteria have to be considered, as is the case of choosing a search engine. A future extension of this study will include the application of MCDA techniques [11] in choosing the most appropriate tool for searching the greek WWW.

References

1. Sim S., Davies J.: *The Internet and Beyon*d, British Telecommunications plc, (1998)
2. Dertouzos M.: *What Will Be: How the New World of Information Will Change Our Lives*, HarperBusiness (1998)
3. Gourvennec Y.: *Information Tracking in the Information Age*, Visionary Marketing, (1999), http://visionarymarketing.com.
4. Kaimaki V.: *Science et Media, Science et Technologies Modernes*, PARIE VII (1997).
5. Sonnenreich W. and MacInta T.: *Guide to Search Engines*, John Wiley (1998)
6. Glossbrenner A.: *Search Engines of the World Wide Web*, Peachpit Press (1999)
7. Veglis A.: Locating Information in Greek on-line News Resources, *Mesogeios*, 16:177-191, (2002)
8. Veglis A.: Communicating with Greek Newspaper via the World Wide Web, *IEEE Global Communications Newsletter*, September (1999) 1-4.
9. Sherman C., Price G.: *The Invisible Web: Uncovering Information Sources*, CyberAge Books, (2001)
10. Lawrence S., Giles C.L.: Searching the World Wide Web, *Science*, 280(5360):98-100, (1998)
11. Bana e Costa C. (ed.), *Readings in Multiple Criteria Decision* Aid, Springer, (1990)
12. Web site, http://www.google.com, accessed April 2003.
13. Web site, http://www.altavista.com, accessed April 2003
14. Web site, http://www.trinity.gr, accessed April 2003.
15. Web site, http://www.phantis.gr, accessed April 2003.
16. Web site, http://www.pop.gr, accessed April 2003.
17. Web site, http://www.anazitisis.gr, accessed April 2003.
18. Web site, http://www.searchengineshowdown.com, accessed April 2003.
19. Web site, http://www.searchenginewatch.com, accessed April 2003.