International Journal of Scientometrics, Informetrics and Bibliometrics
*ISSN 1137-5019*



December 2, 2003



# VOLUME 5 (2001): ISSUE 1. PAPER 1

# The Responsiveness of Search Engine Indexes

**Mike Thelwall**
School of Computing, University of Wolverhampton
Wulfruna Street, Wolverhampton WV1 1SB, UK
m.thelwall@wlv.ac.uk

## Abstract

Search engines are an important tool for information foraging on the web. The broad details of how they work is, therefore, of relevance to both information seekers and providers. Yet search engines are known to only index a fraction of the web, up to a maximum of 16% in one recent study. A search engine must crawl the web periodically in order to maintain an up to date index, but, given the limitations of total coverage, how can it decide which sites to cover and which to ignore? One answer lies in research showing the importance of web links in identifying useful sources of information. This paper reports on an experiment to investigate the effect of link count on the indexing of 1000 sites in three search portals over a period of seven months. It was found that, although all engines added sites during the period of the survey, only Google showed evidence of being very responsive to the existence of links on the test site, whereas AltaVista's results were very stable over time.

## Keywords

search engines, web, link analysis, time series

## Introduction

Search engines are suites of computer programs that automatically find and download web pages, then storing them in a database. They include programs that link the database to a user interface, normally in the form of a web page, so that it can be interrogated through the Internet, often by a keyword search. The size of the web is such that no search engine has been able to maintain a database indexing all or most of it, with 16% of a subset of the web being the maximum found in one February 1999 survey, (Lawrence and Giles, 1999). See Dahn (2000) and the original article, however, for a discussion of the scope and reliability of such estimates. This figure appears to be decreasing compared to a previous study (Lawrence and Giles, 1998), although one search engine has recently claimed a major increase in the proportion of the web indexed (Google, 2000c). An existing search engine can find web pages in one of three ways, see Mauldin (1997) for example.

1. From its list obtained by previous crawls of the web.

2. By extracting the links out of known web pages.

3. From human input, such as the submission of URLs by web site owners.

Logically, any search engine either indexes every URL that it can obtain using the above three methods, or it must have a procedure for determining when not to index a known URL. It is believed that the latter is standard for search engines as a result of the high level of interconnection of a large part of the web (Broder et al., 2000). The issue of the exact workings of this procedure in the major search engines is of vital importance to web site designers that desire their pages to be accessible through search engines, yet these details are likely to be complex for various reasons (Cho et al., 1998). The commercial nature of search engines, however, means that these algorithms are often a proprietary secret. Detailed information is known, nevertheless, about the original Google algorithm (Brin and Page, 1998). It is an iterative algorithm that ranks pages by order of importance by assessing the links pointing to them and the standing of those linking pages, which is also measured by a count of links to them and the standing of the linking pages. Pages may, therefore, not be able to be assigned a page rank or indexed, "because not enough other pages on the web link to them" (Google, 2001). This shows that search engines can use links not only to find pages, but also to decide whether to index known URLs.

Search engine databases are updated over time in response to the changing web, parts of which are very static, whilst others are very dynamic (Brewington and Cybenko, 2000). Time series have been used as a tool to analyse the changes in information availability from search engines (Bar-Ilan, 1999; Mettrop and Nieuwenhuysen, 2000) and their reliability for use in cybermetrics (Rousseau, 1999). Bar-Ilan found that search engines both lost information and forgot URLs, whilst Rousseau discovered that AltaVista was very unreliable, recommending a five point median to smooth out its errors. Rousseau also discussed the fact that search engine algorithms change periodically and so the temporal extrapolation of specific findings about reliability should be treated with caution. A fundamental problem is that the objectives of a commercial search engine designer are unlikely to prioritise factors of importance to academics, such as long term reliability or repeatability, or even comprehensiveness of coverage. A more likely prime objective is to be able to provide the most relevant and useful information to users, with a consequent focus upon the quality of information available (Maudlin, 1997; Cho et al., 1998; Kirsch, 1998; Brin and Page, 1999).

Apart from Google, little appears to be known about the procedures by which search engines create the database space to index new sites, other than the removal of dead pages. What has been observed, however, is an apparent periodic renewing of databases (Rousseau, 1999). Since most continue to offer a facility to add links to the database, although Google, for example, does not guarantee to index added URLs (Google, 2000a), it is presumed that method 3 above is a generally effective mechanism. What is not known, however, is whether in the context of finite resources there are procedures for finding new sites from links (method 2 above), at the expense of old sites (method 1). This is an important issue because the majority of commercial web sites, for example, are not indexed at all in some search engines (Thelwall, 2000a). If it were to be found that method 2 was not used in mature search engines then this would make the manual submission to search engine of the URLs of new sites vital for their designers.

The importance of web links in aiding discovery of the most useful information on the web has been widely recognised in information science (Ingwersen, 1998; Thelwall, 2000b) and computing (Amento et al., 1999; Dean and Henzinger, 1999; Kumar et al., 1999; Rafiei and Mendelzon, 2000), and has motivated the design of Google’s page selection and ranking algorithm. Essentially a page seems more likely to contain useful information if other

pages link to it, and the context of that usefulness may also be indicated by an analysis of the linking pages. This basic idea has been developed successfully in several directions in order to solve different information retrieval tasks. One strand, for example, that is separate from, but related to, search engine design, has looked at the problem of finding the best pages for a chosen topic. Approaches tried include algorithms that follow links from a topic-related starting site and those that attempt to differentiate between pages that contain authoritative information on a subject and ones that are hubs of information sources, linking to many useful sites for a topic (Gibson et al., 1999). The success of these and other approaches provide evidence that the link structure of the web does contain useful information that information providers, such as search engines, would benefit from using. It is, therefore, relevant to explore the extent to which search engines use the links found.

## Methodology

An experiment was devised to test the responsiveness of search engines to information about previously unknown web sites through links to those sites. In order to make the test as realistic as possible, it was decided to use a large collection of real web sites that were not indexed in search engines. These sites therefore had to be found by a method other than through search engines. The procedure used was to create a large number of legal domain names and to test them to see if they were in use. The domains from **www.a.co.uk** to **www.zzzz.co.uk** were tested, with 4,700 found to be active. The UK domain was used because the survey was based in the UK and search engines have some ability to identify the location of web requests, and may choose to add URLs to a regional database rather than the international version. The main commercial domain was selected as a source of large numbers of web sites unknown to search engines (Thelwall, 2000a). Some search engines allow users to directly check whether a site is in its index by a search request for the URL, perhaps using special syntax. Three of the most used search engines possessing this feature were used. Experiments with other leading search engines at the time, including Google, were unable to find a reliable method of diagnosing whether they indexed pages on a site. A thousand of these sites were selected at random to be part of the experiment. Five hundred of the selected sites were chosen, again at random, to be part of the test group and the remaining 500 formed the control group. A web site was constructed with links to the test group but not the control group.

The web site was designed so that the pages with links to the 500 sites would not have to be directly submitted to search engines, but would appear to be well-linked to pages. In order to achieve the former aim, and to avoid having large pages that might be incompletely indexed, the links were spread over 26 pages in alphabetical order and all were linked to by a home page. This site was then placed on a university web site known to be indexed by the chosen search engines. A second web site with only cosmetic changes to the first was placed on a commercial web server, but in order to promote the 26 links pages, five additional web pages linking to them and similar to the main page were placed on separate commercial web servers. All the home pages were then registered in the three search engines and the presence or absence of the 1000 sites, including the 500 control group sites, in the three search engines was checked on a weekly or twice-weekly basis from a computer in the UK academic domain. The checking was achieved with the host: advanced command to search for the domain name in AltaVista but with direct home page URL searches in Yahoo! and HotBot. The AltaVista check, then, was for whether any pages in the site had been indexed, whereas the other two were checks for whether the home page had been indexed. Although it is possible that a search engine indexes some pages on a site, but not its home page, no instances of this were found in the data set. All checks were, in effect, testing whether at least one page on the site was in the search engine database. Data was lost, however, for a two week period when the computer collecting the data crashed. In the middle of the experiment the main commercial web site was removed, with the weekly monitoring continuing.

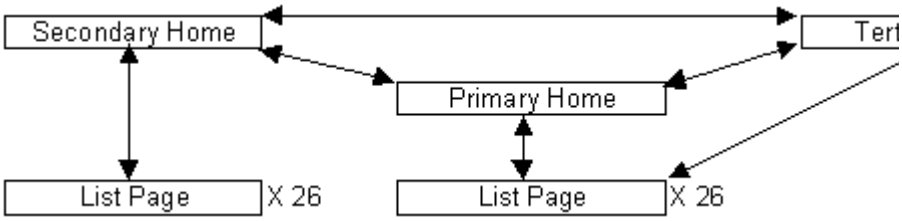


**Figure 1.** The overall design and link structure of the web site.

## Results

The number of web sites indexed by each search engine in the linked group and control group are shown in figures 2 to 4.

**AltaVista**

AltaVista's graph does not show any evidence of having used the links from the test site at all. This is despite evidence from the server logs that its spiders crawled and indexed the site in the first half of the test period, and that referrals of web surfers in response to AltaVista searches had begun to occur by the 7th of June. More generally, its graph was overall by far the most stable of the three plotted. On one date there was a small jump in the number of sites registered, but at all other times the total number of sites changed very little. A more detailed study of the 1000 individual sites showed that the database was not static: sites were being added and others removed almost every week. There were no occurrences of sites being in for only a short period of time, however, and thus, even on the individual site level the database was relatively stable. The 22 sites that had been removed were checked, and it was found that in half of the cases the sites were gone, redirected automatically to another site, set up a frameset with frames from another site, or had banned search engines. Of the remainder, two sites had been changed to contain only graphics or embedded animations, and the common factor with the rest appeared to be the use of complex HTML in various guises, including frames and absolute positioning tags. One relatively rare web editor accounted for three of the sites, suggesting that its design was incompatible with AltaVista. The 35 sites in the test group that were added to AltaVista during the time of the study were also investigated. The advanced search feature of AltaVista that allows a check for pages linking to any given URL was used. It was found that only seven of them did not have any links to them reported in AltaVista, other than links from the test sites. It would be possible, therefore, that most of the sites had been added as a result of following links, but that the links were found a considerable time before they were actually used.

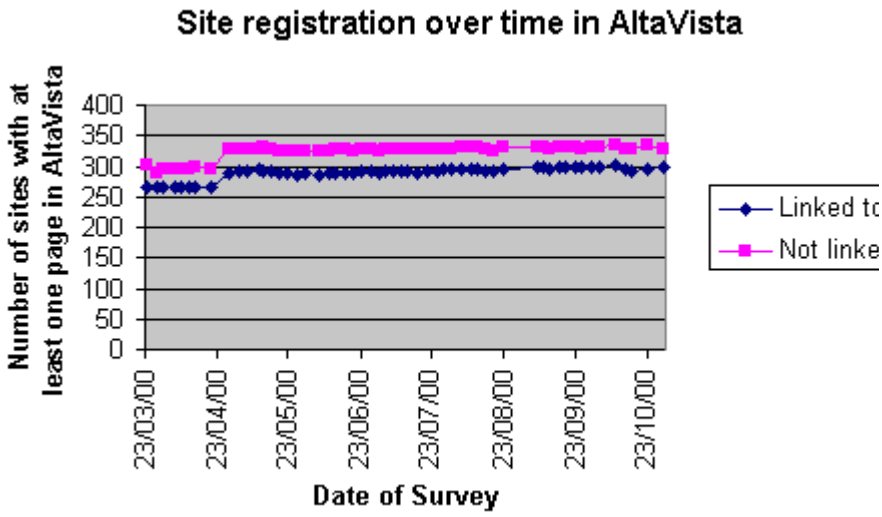


**Figure 2.** Sites with at least one page indexed in AltaVista

**Hotbot**

Hotbot's graph also does not show any evidence of having used the links from the test site. An Inktomi spider did crawl the entire site repeatedly over the period of the study and added links pages to its database, but showed no evidence of having followed the links.

The overall graph showed great variability, including four step jumps and one period of steep climb in site counts. On the level of individual sites, there was even more change. Many sites were being added and excluded every week, and over the months many were included and then excluded repeatedly, each period lasting for weeks or months. In the period from the middle of June to the middle of July, there was a big temporary increase, repeated again at the end of September. It was generally the same sites that were added and removed for the first increase, but there was a partly different set of sites added for the second increase. At the end of the survey period, HotBot changed its method of processing requests for URLs, returning approximate matches when sites were not found. Under this new algorithm, the survey reported would not be possible again with the same level of accuracy.

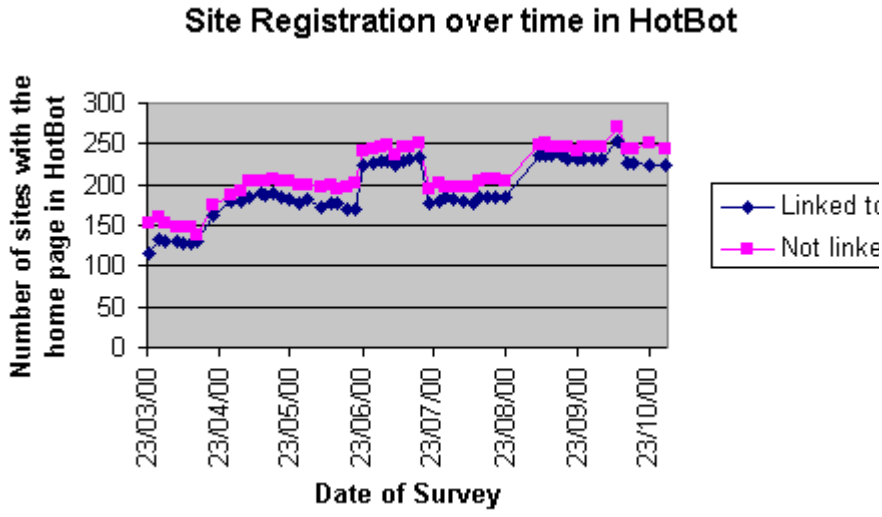


**Figure 3.** Sites with at least the home page indexed in Hotbot

**Yahoo!**

The graph for Yahoo! must be read with caution because of a change in search technology during the test period. Yahoo! is a directory-based service but it is supported by a search engine. This means that any searches not matching main directory entries are passed on to a search engine for matching in the wider web. At the start of the survey, Yahoo! used the same search engine technology as HotBot, Inktomi and their graphs are very similar. A detailed comparison showed that they were not, in fact, identical in results at any one time. It was true, however, that all sites found in one were also found in the other at some stage. This would be consistent with the hypothesis that HotBot and Yahoo! used different versions of the database built with a common algorithm.

The first jump in the graph occurred when Yahoo! switched from using Inktomi for its backing search engine to using Google (Google, 2000b). The remainder of the graph reflects Yahoo!'s use of Google. This causes a problem in that there was consequently no data concerning the registration of the test and control groups of sites in Google before the test site was published. The random nature of the site selection nevertheless allows for statistical tests to show that the difference between the two groups is significant at the 0.01% level. There is, therefore, strong evidence to show that Google has used information from the constructed sites in order to select which pages to index. As can be seen from the graph, however, Google did not index all of the sites in the test group. A selection of the sites not indexed were analysed to discover the reason why, and it was found that whilst about a third had gone or contained holding pages, the rest contained features that would make them difficult to index with search engines, such as framesets, complex HTML, embedded animations and JavaScript based links. Of the 26 sites studied, only one was a plain site that would be easy to index, but it may have been the case that this site was offline when the crawler tested it. The drop on the graph occurred after the main commercial web site was removed, illustrating Google's ability to respond to drops in site 'popularity' as measured by its link count based algorithm. The removal of a large number of sites was, however, surprising in the light of Google's claims to be the "World's Largest Search Engine" (Google, 2000c).

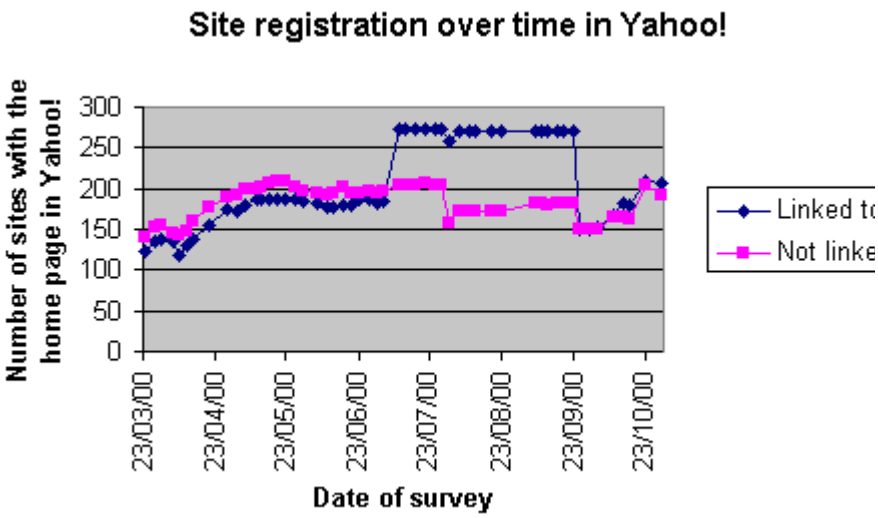


**Figure 4.** Sites with at least the home page indexed in Yahoo!/Inktomi and Yahoo!/Google

## Discussion

The graphs for AltaVista and Hotbot show no significant differences in

behaviour between the test and control groups over time, showing that these search engines, over the time period of the study, did not use the implicit information in the site links in order to build their databases. Both graphs did, however, show significant changes over time, supporting the findings of Rousseau (1999), Bar-Ilan (1999) and Mettrop and Nieuwenhuysen (2000), although AltaVista was much more stable than reported in previous surveys. The graph for Yahoo! does show a clear, and statistically significant, indexing of new sites in the test group, indicating that its search technology at the time, from Google, was able to both find new URLs from the web and to respond to drops in site popularity. Google was reported to be expanding during the early part of the survey, claiming to have created the largest web index (Google, 2000c), and the fact that it managed to add a large number of newly found sites to its index may also have been a result of this, and, therefore, could be a relatively temporary ability. It must not be concluded from this that AltaVista and Hotbot cannot find new URLs from the web because all links came from a single site. It may be the case that both rejected the site as a potential source of new links, or that the links were added to a list to be crawled, but that the list was too long for the test URLs to have yet arrived at the top. It is clear that search engines must have the ability to find new sites by following links in order to build their initial database, and so it would require much more extensive testing to conclude that a search engine, over a set period of time, had not indexed any new sites other than those submitted by site owners.

A general analysis of the results is also appropriate in the context of the differing results and interest in the reliability of search engine indexes from an Information Retrieval perspective. AltaVista stands out as the most reliable over the time period, both in terms of its overall results and in terms of individual web sites. It is, however, not possible to conclude that all types of searches in AltaVista would be similarly reliable or that it will remain stable. Because search engines may maintain simultaneous different databases in order to cope with high volume usage, it could also be the case that the component that deals with advanced searches is more stable than the others, either in terms of the database itself or the database querying software. The results are, however, encouraging. Hotbot/Inktomi and Yahoo!/Google, in contrast, show the size of fluctuations in the control group that could render any observations deceptively stable over short periods of time, whilst very unstable over a period of a month. This is clearly not a desirable characteristic from a researcher's point of view.

These results should not be read as an evaluation of the search engines involved, as such a task would require a much more extensive set of factors to be taken into account (Oppenheim et al., 2000).

## Conclusions

This survey provides evidence that the search engine Google is able to respond quickly to changes in the web and is capable of finding new web sites without relying on URL submission. Google has become an important search engine, as evidenced by its use with Yahoo!, and also with Netscape. A web site designer wishing their pages to be found through search engines should, therefore, attempt to get their site linked to by other sites, preferably ones known to be indexed by Google. This reinforces the findings of Thelwall (2000a).

The survey also shows that other important search engines can be unresponsive to the appearance of new web pages, even if these are linked to by known pages. The evidence does not prove that the only way to get sites indexed in this context is by registering the URL directly with the search engines because it is possible that some aspect of the test site design caused it to be rejected as a source of new URLs. It may also be the case that there is a backlog of URLs to be added, making the time between finding a new URL and having the free disk space to index it longer than seven months. The

secrecy of the algorithms used to determine new URLs creates this uncertainty. The results do, however, provide an incentive to register web sites in major search engines, even if sites are well linked to.

For those engaged in information retrieval for commercial, academic or other reasons the apparent differences between search engines is a reminder that the use of a single search engine does not give access to the entire web. In the case where the information was likely to be on a newer site that is not well linked to then the information retriever is at the mercy of the web site designer's knowledge or decision about whether to register their site in search engines as to whether the information is findable at all. This seems likely to have the greatest impact on those with the least web expertise, perhaps including a sizeable proportion of the many small commercial web sites and their potential customers. For those wishing to use search engines to analyse the web, the apparent increased stability of AltaVista, with its range of advanced commands, is good news.

## References

Amento, B.; Hill, W.; Terveen, L.; Hix, D. & Ju, P. (1999). An empirical evaluation of user interfaces for topic management of web sites. In CHI 99 Conference proceedings, pp. 552-559. New York: Addison Wesley.

Bar-Ilan, J. (1999). **Search engine results over time - a case study on search engine stability**. **Cybermetrics**, 2/3.Paper 1. <http://www.cindoc.csic.es/cybermetrics/v2i1p1.htm>

Brewington, B.E. and Cybenko, G. (2000). Keeping up with the changing Web. **Computer**, 33 (5):52-58.

Brin, S. and Page, L. (1998). The anatomy of a large scale hypertextual web search engine. **Computer Networks and ISDN Systems**, 30: 107-117.

Broder, A., Kumar, R., Maghoull, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000), Graph structure in the web. **Computer Networks**, 33(1-6):309-320.

Cho, J., Garcia-Molina, H. and Page, L. (1998), Efficient crawling through URL ordering, **Computer Networks and ISDN Systems**, 30(1-7):161-72.

Dahn, M. (2000).**Counting Angels on a Pinhead: Critically Interpreting Web Size Estimates**. **Online**, 24(1):35-40. <http://www.onlineinc.com/onlinemag/OL2000/dahn1.html>

Dean, J. and Henzinger, M. R. (1999). Finding related pages in the World Wide Web. **Computer Networks**. 31(11-16):1467-79

Gibson, D., Kleinberg, J. & Raghavan, P. (1998). Inferring web communities from link topology. Hypertext 98: Ninth ACM Conference on Hypertext and Hypermedia. ACM, New York, NY, USA

Google (2000a). <**http://www.google.com/addurl.html**> (Accessed 18 December, 2000).

Google, (2000b). **Yahoo! selects Google as its default**

**search engine provider**.
<http://www.google.com/press/pressrel/pressrelease25.html>
(June 26, Accessed 18 December, 2000)

Google, (2000c). **Google launches world's largest search engine claims almost all of web**
<http://www.google.com/press/pressrel/pressrelease26.html>
(June 26, Accessed 18 December, 2000)

Google, (2001). **Why doesn't Google index any of my pages?** <http://www.google.com/help/faq.html#nonindex>
(January 17, Accessed 12 February, 2001)

Ingwersen, P. (1998). Web Impact Factors. **Journal of Documentation**, 54(2), 236-243.

Kirsch, S. (1998). Infoseek's experiences searching the Internet. **SIGIR Forum**, 32: 3-7.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tompkins A. (1999), Trawling the web for emerging cyber-communities. **Computer Networks**, 31(11-16):1481-93.

Lawrence, S. and Giles, C. L. (1998), Searching the World Wide Web. **Science**, 280: 98-100.

Lawrence, S. and Giles, C. L. (1999), Accessibility of information on the web. **Nature**, 400:107-109

Mauldin, M. L. (1997). Lycos: design choices in an Internet search service. **IEEE Expert**, Vol. 12(1): 8-11.

Mettrop, W. and Nieuwenhuysen, P. (2000). The reliability of Internet search engines: fluctuations in document accessibility. Proceedings of the 21st National Online Meeting, New York, pp.271-80.

Oppenheim, C., Morris, A. and McKnight, C. (2000), The evaluation of WWW search engines. **Journal of Documentation**, 56(2):190-211.

Rafiei, D. and Mendelzon, A. O. (2000), What is this page known for? Computing Web page reputations. **Computer Networks**, 33(1-6):823-835.

Rousseau, R., (1999). **Daily time series of common single word searches in AltaVista and NorthernLight**.
**Cybermetrics**, Vol 2/3, Paper 2.
<http://www.cindoc.csic.es/cybermetrics/v2i1p2.htm>

Thelwall, M. (2000a). Commercial web sites: lost in cyberspace?. **Internet Research**, 10(2):150-159.

Thelwall, M. (2000b). Extracting macroscopic information from web links. University of Wolverhampton.

Received 20/December/2000
Accepted 22/February/2001