Data Mining through Internet Search Engines: The Case of for Islamic Management Materials

Suhaimi bin Mhd Sarif[1], Yusof bin Ismail[2]

[1]Department of Business Administration
Kulliyyah of Economics and Management Sciences,
International Islamic University Malaysia (IIUM),
Batu 8 ¼ Jalan Gombak
53100 Kuala Lumpur, MALAYSIA
suhaimims@iiu.edu.my

[2]Department of Business Administration
Kulliyyah of Economics and Management Sciences,
International Islamic University Malaysia,
Jalan Gombak
P.O. Box 10, 50278 Kuala Lumpur MALAYSIA
yusof@iium.edu.my

ABSTRACT

The study is carried out to ascertain relative availability of the materials of Islamic management using data mining through the Internet search engines (ISE's). The contents of thirteen ISE's are analysed using a specially created format. The present study established the relative ratios of the coverage of selected terms in terms of number of hits shown within each ISE. The ratios are obtained by dividing the hits for a search term into the means (average) of comparable hits as the standardized denominator. The results of the analysis show that the ISE's contain scarce amount of materials on Islamic management.

Keywords
Data Mining, Internet Search Engine, Islamic Management

1.0 INTRODUCTION

The Internet has been instrumental in providing researchers with instant information on diverse topics on Islam. Although abundant literature is still preserved in hard copy, or print format, one cannot deny the contribution of the Internet as the fastest mean of accessing information. Due to the increasing popularity and importance of the Internet as a source of literature search, perhaps also on Islam, it is crucial to ascertain apparent availability of materials on Islam in the electronic form via the Internet.

The study is carried out primarily to identify the relative frequency of Islamic management materials via the Internet search engines. However, it is outside the scope of the paper to investigate the quality of the contents of the Web-based materials available. The following combination of search terms were used in the study: religion, Islam, Christianity, Muslim, Christian, management, organization, organisation, Islam and management, Islam and organization, Islam and organisation, Muslim and management, and Christian and management.

2.0 LITERATURE REVIEW

The study used Internet Search Engines approach to data mine Islamic Management materials. According to Han and Kamber (2001, p.7), data mining is "*the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories*." As for Witten and Frank (2000, p.9), data mining is "*the extraction of implicit, previously unknown, and potentially useful information from data*." Based on the two definitions, any databases or data warehouses including Internet Search Engines can be used to explore the availability of data, information and knowledge on Islamic Management.

Internet Search Engines (ISEs) are able to perform data mining. Mettrop and Nieuwenhuysen (2001) evaluated thirteen Internet search engines, namely AltaVista,

EuroFerret, Excite, HotBot, InfoSeek, Lycos, MSN, NorthernLight, Snap, WebCrawler, Ilse, Search.nl, and Vindex. They found that those ISEs are suitable for data mining when they showed complications with the behavior and performance of the Internet search engines. The fluctuations in the search results affect the efficiency of Internet search as a publication/communication medium.

Dreilinger and Howe (1997) found that large data obtained from the ISE's may pose difficulties in the selection process. It is also time consuming to filter them, which eventually may not certainly be useful. Gordon and Pathak (1992) evaluated eight ISEs and found that information retrieved from one particular ISE can also be retrieved from a few ISEs. This indicates that the enormous results retrieved may be duplicative, and confusing. Tomaiuolo and Packer (1996) found that similar results in one ISE also appeared in the other ISEs. The repetition can be used as checker against the accuracy of the information. Lake (1997) is concerned about quantity of data than the value (content) of the information retrieved.

Leighton and Srivastana (1997) argue that such rich results give more confidence in terms of information accuracy, not so much of effectiveness of information retrieval. Wang, Xie and Goh (1999) contend that search engines are widely used as tools to find useful information from the Internet. However, most search engines were developed on the basis of technical requirements and without much consideration for the customer's perspective. Ideally, ISE should be very helpful not only to the designers, but also to the users.

Sullivan (2010) identified fourteen Internet Search Engines (ISE's) which he classified as top choices (Google, Yahoo, and Ask), strongly considered (Alltheweb, AOL, HotBot, and Teoma), and other choices (Altavista, Gigablast, LookSmart, Lycos, MSN, Netscape, and Open Directory).

3.0 METHODOLOGY

The study aims to ascertain the number of hits of the materials on Islamic management that emerged from the Internet search engines (ISE's). Major search terms used include religion, Islam, Christianity, Muslim, Christian, management, organization, organisation, Islam and management, Islam and organization, Islam and organisation, Muslim and management, and Christian and management. Related terms are included in the search in order to generate comparative data. For instance, how many hits are produced when the term "Islam," is used in comparison with "Christianity"?

The Internet Search Engines used in the study comprise those verified in Sullivan's (2010) study: top choices (Google, Yahoo, and Ask), strongly considered (Alltheweb, AOL, HotBot, and Teoma), and other choices (Altavista, Gigablast, LookSmart, Lycos, MSN, Netscape, and Open Directory).

The Internet was accessed on 15 January 2010 (Friday) from 4.15 p.m. to 4.35 p.m., and 5.40 p.m. to 6.00 p.m; and 16 January 2010 (Saturday), from 9.25 a.m. – 9.55 a.m. Kuala Lumpur time. The total time spent for the search is 70 minutes.

A tabular format is created to capture the data of interest for each search engine. The ratios for number of hits within each ISE are computed by dividing the hits by the average hits for the search engine. The ISE's mean (average) is used as the denominator for the hits of each selected search term.

The use of ratios to compute hits produced by the Internet search engines provides a means to assess impacts (Bharat & Broder, 1998). Moreover, it is more appropriate to make inference from the use of ratios (Lawrence & Giles, 1998), although such approach can be less reliable statistically (Chu & Rosenthal, 1996). However, given time constraints, it is still economical to use ratio approach (Lawrence & Giles, 1999).

There are five stages involved in the process of computing the ratios:

Stage 1: Identifying search engines
Stage 2: Recording number of hits
Stage 3: Computing the ratios
Stage 4: Consolidating all the ratios
Stage 5: Performing the analysis

4.0 FINDINGS

The results of the study are presented according to the five steps enumerated above.

4.1 Stage 1: Identifying the Search Engines

The study adopted the list of ISE's identified by Sullivan (2010). Table 1 shows the search

engines used in the study to evaluate the presence of specified search terms.

Table 1: Internet search engines browsed.

Top Choices	Strongly considered	Other choices
Google	AllTheWeb.com	Altavista
Yahoo	AOL Search	Gigablast
Ask	HotBot	LookSmart
		Lycos
		MSN Search
		Netscape Search
		Open Directory

4.2 Stage 2: Recording the Number of Hits

The researchers recorded the number of hits shown for all the search terms. The results of the search are shown in Table 2A.

Table 2A: Number of hits for specified search terms.

Search terms	Google	Yahoo
Religion	518,000,000	313,000,000
Islam	132,000,000	70,900,000
Christianity	65,500,000	47,700,000
Muslim	110,000,000	66,600,000
Christian	550,000,000	356,000,000
Management	3,470,000,000	1,330,000,000
Organization	1,890,000,000	452,000,000
Organisation	469,000,000	200,000,000
Islam and Management	16,000,000	6,400,000
Islam and Organization	20,900,000	8,990,000
Islam and Organisation	7,280,000	3,710,000
Muslim and Management	15,400,000	5,820,000

Search terms	**Ask**	**AlltheWeb**
Religion	69,300,000	240,000,000
Islam	8,735,000	47,700,000
Christianity	11,550,000	43,100,000
Muslim	10,550,000	55,500,000
Christian	73,920,000	270,000,000
Management	202,060,000	1,110,000,000
Organization	80,150,000	427,000,000
Organisation	20,930,000	96,400,000
Islam and Management	1,492,000	4,540,000
Islam and Organization	1,441,000	7,740,000
Islam and Organisation	316,900	2,010,000
Muslim and Management	1,257,000	4,880,000

The above Table 2A shows the number of hits produced by four ISE's: Google, Yahoo, Ask, and AllTheWeb. The ISEs show the highest hits for the term 'management'. However, the ISE's exhibit shrinking hits for the term 'Islam and Management'. Out of the four ISEs, Google registered the highest hits for the terms that are combined with "Islam," such as 'Islam and Organization,' 'Islam and Organisation,' and 'Muslim and Management.'

Table 2B shows the number of hits for specified search terms from another five ISE's: AOL Search, Hotbot, Altavista, and Gigablast. All of the ISEs show high hits for the term 'management', with the highest statistic produced by Altavista. Altavista registered the highest hits out of the five ISEs for the term 'Islam and Management'. It is also Altavista that maintains the highest hits for the other terms that are combined with "Islam," such as 'Islam and Organization,' 'Islam and Organisation,' and 'Muslim and Management.'

Table 2B:Number of hits for specified search terms.

Search terms	**AOL Search**	**HotBot**
Religion	3,720,001	46,823,296
Islam	826,667	10,928,097
Christianity	481,334	8,196,514
Muslim	860,001	10,096,560
Christian	3,580,001	80,766,679
Management	22,600,001	180,246,494
Organization	11,866,667	73,406,305
Organisation	3,033,334	30,959,762
Islam and Management	108,667	998,281
Islam and Organization	128,667	1,613,400
Islam and Organisation	47,534	565,450
Muslim and Management	99,334	1,071,615
Christian and Management	606,001	2,820,833

Search terms	**Altavista**	**Gigablast**
Religion	308,000,000	31,323,860
Islam	72,500,000	6,378,361
Christianity	48,100,000	4,849,666
Muslim	63,900,000	4,292,502
Christian	341,000,000	34,447,507
Management	1,260,000,000	77,430,090
Organization	453,000,000	25,264,130
Organisation	201,000,000	11,010,050
Islam and Management	6,400,000	318,665
Islam and Organization	9,010,000	1,309,518
Islam and Organisation	3,710,000	131,695
Muslim and Management	5,870,000	27,291
Christian and Management	32,900,000	34,450,092

4.3 Stage 3: Computing the Ratios

The computation of ratios excluded search engines LookSmart because it produced no hits

when accessed. The ratio for each search engine is computed by dividing the hits by the mean of hits for the search engine. The higher the ratio the higher the relative coverage of the search terms in the search engine.

The terms 'Islam' and 'Muslim' in Google, Yahoo, and Ask produced lower ratios, ranging from 0.001 to 0.003. As for AlltheWeb, AOL Search, and Hotbot, they also generated lower ratios, ranging from 0.001 to 0.004. In Altavista, and Gigablast, Lycos and MSN Search, also lower ratios, ranging from 0.000 to 0.006. LookSmart is excluded in the analysis due to its inability to produce any hits. In Netscape Search and Open Directory, they produced lower ratios: 0.001 and 0.003 for Netscape Search, and 0.000 and 0.001 for Open Directory.

4.4 Stage 4: Consolidating All the Ratios

All the ratios of hits for specified search terms on Google, Yahoo, Ask, Alltheweb, AOL, Hotbot, and Teoma produced lower ratios. For instance, individual term 'Management' produced 0.472, but when it combined with Islam as 'Islam and Management', it produced 0.002. Likewise, when 'Organization' combined with 'Islam', it produced 0.001. This trend shows the search engines could produce more hits for individual terms, but when it combined with another term, it produces lower hits.

4.5 Stage 5: Performing the Analysis

The last stage of the analysis identified the highest ratios generated by each search engine. All the ISE's could capture the terms "Management," "Organization," and "Organization," however, they produced comparatively very low ratios when they were combined with "Islam." For example, in Google, the term "Management" shows 0.472, but for "Islam and Management" the ratio declined to as low as 0.002. Learners and researchers may not consider the Internet as a reliable source for learning about Islamic management.

5.0 DISCUSSION

The study found that 12 out of 13 Internet Search Engines (ISE's) used are capable to search materials for Islamic management except 'LookSmart' search engine. High number of hits for search terms are produced for terms that do not contain 'Islam' or 'Muslim.' When they are combined with these terms, the number of hits and ratios dropped immediately. This trend shows that materials on Islamic management or 'Islam' and 'Muslim' are really scarce in the Internet Search Engines.

The study constitutes a preliminary attempt to use Internet Search Engines for knowledge data mining on the availability of Islamic management materials. The relative ratios were computed solely for each search engine, and could not be compared in a robust manner without ascertaining the common denominators for all the search engines combined. In order to obtain more valid ratios, both the numerators and denominators of all the search engines in the set need to be corrected by some statistical procedures.

The data used for the study is cross-sectional. A meaningful pattern would emerge if the data used is longitudinal, at least over a few quarters within a year.

6.0 CONCLUSION

Based on the ISE results on the specified search terms, data mining approach via ISE is capable exploring the study of Islamic Management although the amount of materials on Islamic management is really scarce. This suggests any research who is attempting to find materials on Islamic management from the Internet Search Engines will face difficulties to write literature on Islamic management.

It should be noted that the existence of the terms in the ISE's in no way suggesting that the adequacy and quality of the contents of the materials.

REFERENCES

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. Paper presented at the *7th International World Wide Web Conference*, April.

Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. *ASIS*. From http://www.asis.org/annual-96/Electronic-Proceedings/chu.htm

Dreilinger, D., & Howe, A. E. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems, 15*, 195-222.

Foxall, G. R., & Fawn, J. R. (1992). An evolutionary model of technological innovation as a strategic management process. *Technovation, 12*(3), 191-202.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann.

Ismail, Y., & Sarif, S. M. (2006). The coverage of maqasid al-shari'ah in the Internet search engines. Paper presented at the *International Conference on Islamic Jurisprudence and the Challenges of the 21st Century*, Kuala Lumpur, 8-10 August.

Lake, M. (1997). *2nd Annual search engine shoot-out* [Electronic Version]. *PC Computing* http://www4.zdnet.com

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

Lawrence, S., & Giles, C. L. (1999). *Accessibility and distribution of information on the web. Nature*, 400, 107-110.

Leighton, H. V., & Srivastava, K. (1997). *Precision among World Wide Web search engines (*search engines): AltaVista, Excite, HotBot, InfoSeek, Lycos. . In.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-6251.

Sullivan, D. (2010). *Major search engines and directories*. Retrieved 15 January, from http://searchenginewatch.com/author/index.php/DannySullivan

Tomaiuolo, N. G., & Packer, J. G. (1996). An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries, 16*(6), 58-62.

Wang, H., Xie, M., & Goh, T. (1999). Service quality of internet search engines. *Journal of Information Science, 25*(6), 499-507.

Witten, I. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco, CA: Morgan Kaufmann.