San Jose State University

From the SelectedWorks of G. Kent Webb

2016

INTERNET SEARCH ENGINE CAPTURE SUCCESS RATES AND MORTALITY STATISTCS FOR HYPERLINKS TO PUBLIC NEWS ARTICES

G. Kent Webb

This work is licensed under a Creative Commons CC_BY International License.

Available at: https://works.bepress.com/kent_webb/32/

Issues in Information Systems
Volume 17, Issue II, pp. 25-33, 2016

INTERNET SEARCH ENGINE CAPTURE SUCCESS RATES AND MORTALITY STATISTICS FOR HYPERLINKS TO PUBLIC NEWS ARTICLES

G. Kent Webb, San Jose State University, g.webb@sjsu.edu

ABSTRACT

As an input to an online knowledge base supporting public deer management, the internet was searched for relevant information on a daily basis for over 2,000 days. Although an intensive search relying on about 60 keywords and phrases was employed using Google Alerts, a simple additional search using Microsoft Bing found a significant number of articles not discovered by Google Alerts, about 14 percent of the total. This suggests an important advantage in using more than one search engine to insure a more thorough search. A sample of the information from each source captured by the search process was stored in a knowledge base available on the internet, subject to copyright restrictions, with a hyperlink to the original article. One drawback to this approach was the number of dead links, hyperlinks that no longer worked because the articles were moved or removed, reducing easy access to the original article. The percentage of dead links rose over time with links between 1 and 200 days old having a mortality rate of 4.8 percent. Links between 1801 and 2000 days old had a mortality rate of 47.2 percent.

Keywords: Knowledge Management, Internet Search, Copyright, Environmental Scanning

INTRODUCTION

The internet provides a vast amount of information that can be used to assist in decision making. Search results routinely report finding millions of relevant search results on many topics, but most search engines limit access to the first several hundred results. As an ongoing project to create a knowledge base supporting public deer management, the internet has been searched on a daily basis since mid-2010. The process of searching the internet for relevant information is one tool of environmental scanning, often used to enable knowledge discovery for business and other organizations (Choo, 1993; Lau, Liao, Wong & Chiu, 2012; Liu, Turban & Lee, 2000; Perry, Taylor & Doerfel, 2003; Reinhold, Wagner & Scholz, 2005; Webb, 2016).

The search engines in this case were run on a daily schedule and restricted to results that were posted within the past 24 hours with the goal of finding every useful piece of information that might help in making decisions. The results were stored by location, topic, and trending issue in a searchable public website -- www.deerfriendly.com -- so that the information could be more easily be discovered and used by the public in the decision making process. In keeping with copyright, since the information is provided in a public web site, only a sample of the most important information is preserved and a link is provided to the source of the information. Some significant articles were stored offline for reference, but the goal of the project has been to make as much information as possible available online. Less than 0.1 percent of the annual users, approaching 200,000, contact the site for more information.

This paper explores two problems that have emerged with the approach. First, although a very thorough search was being conducted using Google Alerts to capture new information, using about 60 key words or phrases, it became clear from doing a few searches using Microsoft Bing that even this very detailed search was failing to capture a significant number of important links to relevant information. Second, the links provided to the original source of information became inactive, or broken – dead links – when the original location of the information was moved or the information was removed. The use of public knowledge bases for environmental management has grown as access to the internet and development tools have become more readily available. For example, the UK Environmental Change Network Data Centre captures and manages data made publically available using a database and Geographical Information System (GIS) (Rennie, 2016).

Reliance on a Single Search Engine

A 2012 study on the use of internet search engines to obtain important medical information concludes that although search engine results highly overlap, users should rely on multiple search engines in order to avoid missing important information (Wang, Wang, Wang, Liang, & Xu, 2012). In practice, people often rely on a preferred search engine. As an example, the data in Table 1 reports the results of a survey of 129 general business majors when they were asked if when doing an important search they use more than one search engine to make sure they are getting all the results. About a quarter of the students reported that they did.

Table 1. Survey of Student Search Engine Reliance

Survey Question: When you are doing an important search, do you often try more than one search engine to make sure you are getting all the results?		
Response	***Number***	***Percent***
Yes	31	24 %
No	98	76 %
Total	129	100 %

Another common practice of internet search is to look at only the top results, sometimes just the first page. A 2003 study of participants looking for medical information on the internet found many selected information from the first page of search results (Peterson, Aslani & Williams, 2003). As a result, improving page rank is an issue for website managers (Baye, De los Santos, & Wildenbeest, 2016). A 2006 study of four search engines -- MSN Search, Google, Yahoo! and Ask Jeeves – examining the results of what appears on the first page of results for similar queries found that "… the percent of total results unique to only one of the four Web search engines was 84.9%, shared by two of the three Web search engines was 11.4%, shared by three of the Web search engines was 2.6%, and shared by all four Web search engines was 1.1%" (Spink, Jansen, Blakely & Koshman, 2006). A 2005 study of page rankings finds considerable differences for the results appearing on the first page among major search engines (Bar-IIan, 2005). A 2012 study looking at the top five results for searches on topics important to children found that both Bing and Yahoo!kids had a nearly 30 percent overlap with Google, although with different rankings (Bilal, 2012).

A simple Google search on the word "deer", the topic of this knowledge base, recently yielded 242,000,000 results. Only about the first 900 results were actually available. More detailed keywords would presumably be required to access the other approximately 241,999,100 results. To get around this problem, searches for this knowledge base used settings available in Google Alerts that potentially allow the user to retrieve automatically every posting on the public internet as it becomes available. In addition to the search phrase and time, there is a setting for the result type with the option, "Only the best results", or the option used in this study, "Everything." While Google Alerts is a very useful search tool, the results of this study suggest that the option "Everything" should not be taken literally and should perhaps be renamed "Everything that Google can find".

Link Mortality, Dead Links

For this knowledge base, a sample of what was determined by manual inspection of each article to be the most important information was copied to the public web site. Although the search engines were good at finding relevant phrases within the article, the most important information relevant to the knowledge base was often missed. The hyperlink to the original article was preserved as documentation for the information so that users of the knowledge base could refer to the original article. For articles with significantly more information than could be publically preserved given copyright restrictions, a copy of the original article was preserved in a database not available online. Hyperlinks preserved on the site included the original title of the article, the publication date, and the publisher. As

a result, even if the hyperlink were no longer active – dead -- many of the original articles should be retrievable by contacting the publisher.

A 2004 study examines the growing use of hyperlinks to interconnect online news, concluding that "linking structure apparently encourages those who commonly use the Web to have more densely interconnected knowledge structures for public affairs topics" (Eveland, Marton & Seo, 2004). Himelboim (2010) takes a network approach to examining the use of hyperlinks in foreign news services to evaluate the flow of information from news sources. The issue of finding and storing available information in the rapidly growing body of information available on the internet using hyperlinks is specifically addressed by Aminu (2015), who develops a mechanism for detecting dead links related to news information in an ontology knowledge base. The motivation for his study was the development of a knowledge base using hyperlinks to information on the internet, similar to the approach described in this paper, because of the vast amount of unstructured information. He concludes that "alternative information-organization approaches are required to more effectively and efficiently navigate and retrieve information from systems, web included…" Morishima, Nakamizo, Iida, Sugimoto and Kitagawa (2009) provide an approach to identifying and finding the new location of links that are dead because they have been moved. In order to support the "web of data", Volz, Bizer, Gaedke and Kobilarov (2009) also provide a toolkit for discovering dead links and maintaining continuous coverage when data links change.

Because news articles are consider to be contributors to the public record, the copyright restrictions for news material are somewhat more liberal (Lopez-Taruella, 2012) than for material that might appear on a commercial site devoted to, for example, product or company information. In a thorough analysis of copyright for archival purposes, a 2003 study commissioned by the Library of Congress concludes that fair use of works is "generally broader for fact-based works" and that "Certain uses are favored in the statute; they include … news reporting … scholarship, and research …" (Besek, 2003). .

A 10 percent rule for the amount taken from the original article has become a common standard (Zaharoff, 2001), although this rule requires other considerations. A good practical example comes from the Las Vegas Sun and other publications owned by Greenspun Media group that has become so concerned about the negative effects of copyright violations on the viability of newspapers that it has employed people to track down and sue copyright infringers exceeding this basic guideline "allowing links and text totaling as much as 100 words or 10 percent of the story, whichever is less" (Green, 2010). This "100 words or 10 percent" rule was used as the maximum sample taken and stored online for this knowledge base. Over the years, many newspaper reporters have contacted us about the using of the site as a source for their own research. All considered that the sampling approach used in this knowledge base to be appropriate. By providing a link to the original article, the knowledge base drives some traffic to the original site and, like a Google search, thereby increases traffic and their advertising revenue.

Copyright rules for online sampling of news are far from finalized and vary significantly in other countries. For example, Google News provides a somewhat similar service by sampling news, providing links, and organizing it by topic or region. At the beginning of 2015, Google News shut down operations for Spain because of a strict new copyright law requiring payment of fees for use of even headlines with links. In 2007 a Belgian court fined Google for displaying links to Belgian newspapers (Sterling, 2014) and decided that Google could not use any exemptions, such as claiming "fair use" (Ott, 2008). News aggregation such as by companies like Google News and the Huffington Post have been blamed by Rupert Murdock of the News Corporation and Dean Singleton of Associated Press as contributing to the decline of traditional media (Isbell, 2010). Isbell's article provides a survey of related copyright cases and best practices for sampling from news sources and discusses the important issue of whether the sample is transformative. The knowledge base considered in this paper would fall under the category of "Specialty Aggregators" which provide readers the service of collecting information of a particular topic in one place and thereby contribute something new and useful.

RESEARCH METHODOLOGY

The analysis presented here focuses on information in public news sources, but the search also includes You Tube videos and scholarly articles found using Google Scholar. Although this knowledge base relies on manual analysis and categorization of captured links, Shen, Wang, Luo and Wang (2013) describe a graph-based framework to create a knowledge base from the 400 million tweets posted on twitter each day. One advantage of the manual search is that the site manager and author of this paper has been compelled to read many articles and other information that has greatly increased his understanding of the important topics. A famous article with nearly 6,000 citations suggests that 10,000 hours of effort are required to become an expert (Ericsson, Krampe, Tesch-romer, 1993).

Figure 1 illustrates the daily search and capture process used to populate the knowledge base. Daily Google Alerts on keywords or phrases such as "deer," "Pennsylvania deer", "deer management," and "Chronic Wasting Disease" are examined and a sample of each relevant article with the hyperlink is posted on the home page of the website so that users can see the information as it arrives. A little more than a week of news is typically enough to fill this page. This page can be used as a daily briefing by users actively involved in deer management.

Using Google trends to examine how people search for information on this topic, it appeared that most users were interested in news by state and a few other trending topics such as "urban management", "populations trends" and "disease of deer." Since deer are managed by each state or at the municipal level, information from the knowledge base is typically used by people involved in decision making at these governmental levels. For example the page providing information on "urban deer management" is the fifth most popular page on the site and has received about 25,000 page views over the course of the project.

After capturing the article and storing it on the home page, the article is also placed on the state page and may also be placed on a trending topics page (Sort and Store, Stage 1 in Figure 1). As these pages become filled, older articles are saved in an archive for each state by major topic, and important articles related to key topics are also archived as subpages of trending topics pages. As a result, a few key articles may be stored more than once, making it easier to find information simply by browsing the site. The site can also be searched using a Google search tool.

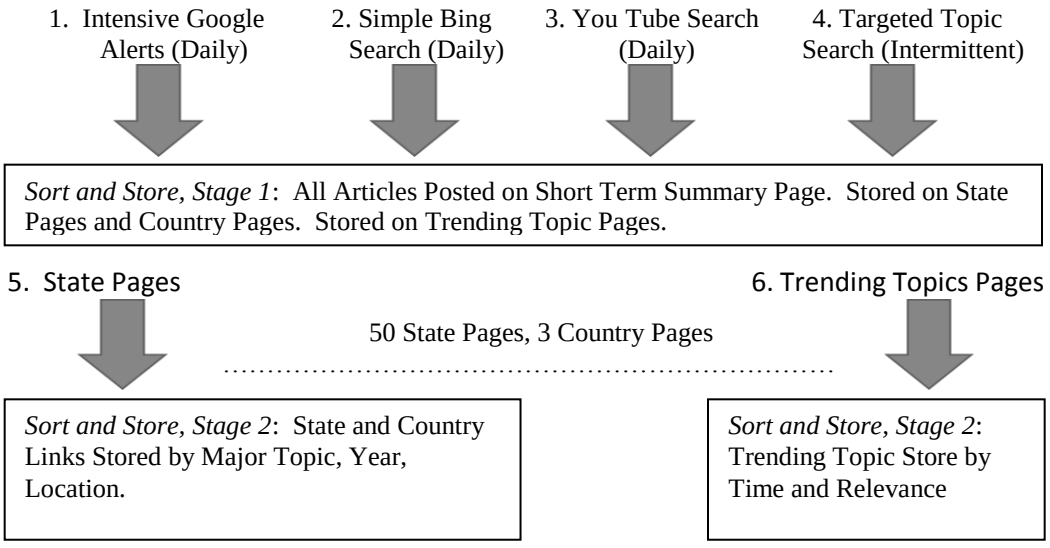

Figure 1. Capture and Sorting Process for Hyperlinks to Public News and Other Information

The two issues related to this process that were analyzed in this paper are, first, what percentage of relevant news articles would have been missed if only the very thorough Google search tools was used. In the spring of 2016, a

sample of 400 news articles posted on the site were tracked based on how they were discovered: both by a simple, daily Microsoft Bing search on the keyword "deer", and by the very thorough, daily Google Alerts search using about 60 keywords and phrases. The second issue examined was how the percentage of dead links increased over time. For this second issue, 1,200 links captured from October, 2010, through March, 2016, were selected in a stratified random sample and tested to see if each link was still active or if the link was dead. The two issues stated as alternatives to the null research hypotheses are:

H_{A1}: Even a thorough search by a single search engine will miss a significant number of relevant results that will be captured by using a second search engine.

H_{A2}: Link mortality, the number of dead links (inactive links), will increase with time.

RESULTS

Search Engine Capture Results

Table 2 presents the results of the search and capture process, reporting the number and percentage of total news articles found over a sample of about 45 days by search engines Google and Bing. The Google search was an intensive daily search, using Google Alerts on about 60 keywords or phrases related to deer management restricted to the previous 24 hours. The Bing results were from a simple search on the keyword "deer" restricted to news articles over the past 24 hours. Since 5 percent is commonly used as a measure of significance, the null hypothesis tested as presented in Table 2 is that 5 percent or less of total relevant articles that could be captured by a search engine would be missed by Google Alerts.

Table 2. Number and Percent of News Articles Captured by Google Alerts and/or Bing Daily Searches. Test of Hypothesis 1

Search Engine	***Number of Articles***	***Percent of Total Articles***
Google and Bing	139	34.75 %
Google Only	206	51.50 %
Bing Only	55	13.75%
Total	400	100.00 %
Testing the null hypothesis that the percent of news articles missed by Google Alerts is less than or equal to 5 percent of the total articles.		
*Hypothesis **	*Significance***	*Conclusion*
H_o: $p \leq 0.05$	0.000	Reject the null
* where p is the proportion of articles missed by Google Alerts and discovered by Microsoft Bing ** calculated using one-sample binomial test with SPSS 22, significance rounded to three decimal points to the left of the decimal		

As reported in Table 2, a one-sample binomial test of the null hypothesis that a thorough Google Alerts search will miss only five percent or fewer of available news articles is rejected with high confidence. Google Alerts missed 13.75 percent of total articles discovered when a simple daily search by Microsoft Bing was added to the process. A simple search on one key word using Microsoft Bing captured 48.5 percent of relevant news articles (the total of Google and Bing), suggesting that users relying on even a daily search are missing about half of relevant information sources by using only one search engine and a single keyword.

Link Mortality Statistics

Links to 1,200 news articles posted to the knowledge base were selected using a stratified random sample and evaluated during early spring, 2016, and coded as active or inactive (dead). The number of days since the hyperlink

had been posted was calculated. A binary logistic regression model was used to evaluate the probability that a link would become inactive over time for the links that had been posted from 1 to 2,000 days since the evaluation period. Table 3 presents the results, indicating that link mortality increases over time. Variable significance is calculated to three significant decimal places. The Wald chi-square test is used to test the hypothesis that B coefficient for the constant and the Days variable is equal to zero. In this case, the null hypothesis is rejected with a high level of significance. The column labeled "Exp(B)" is the exponential of the B coefficient, an odds ratio.

Table 3. Binary Logistic Regression Analysis of Hyperlink Mortality, the Percent of Dead Links Based on Number of Days Since the Hyperlink Was First Posted. Test of Hypothesis 2

Variables	***B***	***Wald***	***Significance****	***Exp(B)***
Constant	-2.531	117.3	0.000	1.001
Days Since Posting	0.001	205.159	0.000	0.080
* Significance calculated to three decimal places. Null hypothesis rejected with significance below 0.000. Dead links were coded as "1", active links as "0"				

Table 4 presents the data for the percent of dead links observed in the sample, organized in categories of days since posting with a range of 200 days for each category. The predicted number for the midpoint of the "Days Since Posting" category from the binary logistic regression model are also presented. The predicted percent is high for the lower number of days and high for the highest number days. Some online news sites have adopted the policy of keeping online news articles in place for a long period of time since the cost of online storage is low and persons interested in using the news for historical information may still generate some revenue. Online ads are typically automatically updated by web software.

Table 4. Percent of Dead Links Categorized by Number of Days Since Posting on the Internet

Days Since Posting	**Observed Percent of Dead Links**	**Predicted Percent**
1 to 200	4.84 %	8.4 %
201 to 400	10.53 %	10.9 %
401 to 600	7.97 %	14.1 %
601 to 800	23.48 %	17.8 %
801 to 1,000	24.58 %	22.4 %
1,001 to 1200	33.06 %	27.7 %
1,201 to 1400	38.05 %	33.7 %
1,401 to 1600	39.00 %	40.4 %
1,601 to 1,800	48.87 %	47.3 %
1,801 to 2,000	47.20 %	54.4 %

Figure 2 illustrates the observed number of dead links, the link mortality, and the number predicted for the mid-point of each category for the "Number of Days" since the link was posted using the binary logistic regression.

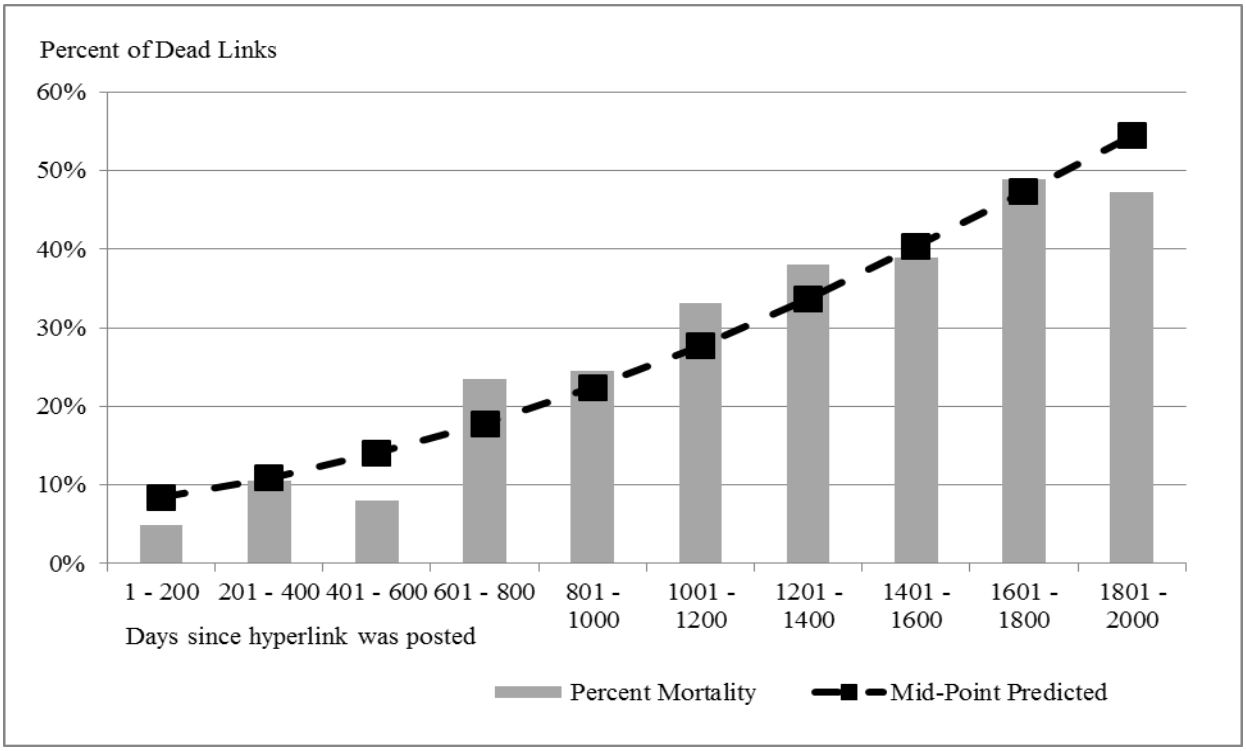

Figure 2. Mortality Trends for Hyperlinks, the Actual Percent of Dead Links Based on the Number of Days Since the Link Was First Posted on the Internet and the Percent Predicted for the Mid-point of Each Category for the Number of Days by the Binary Logistic Regression. Data Grouped in 200 Day Categories

CONCLUSIONS

Even a relatively thorough internet search procedure using a single search engine may miss a large number of relevant information sources. A daily internet search using Bing and a single key word missed 51.5 percent of relevant news articles discovered by a more thorough daily search using Google Alerts. Even the thorough, daily search using Google Alerts missed 13.75 percent of the news articles when results from Bing were included. Although the specifics of search algorithm details are proprietary, there appears to be enough difference that use of multiple search engines will reduce the risk of missing relevant information.

The use of hyperlinks to build the interconnected web creates a management problem for dead links to external sources. In this study, nearly half of the links were no longer active after about 1600 days. Web site managers facing this problem have some online tools that are available to identify dead links. Much of the work on automating the process of finding links that are still available, but have been moved, is mostly in the research stage at this point. In many cases, the information has simply been removed from public access on the internet so the information is lost to researchers relying on internet search. One goal of the knowledge base being developed for this research project has been the preservation of information that will assist in public decision making. Archival efforts face copyright restrictions that are changing and becoming very restrictive in some countries.

REFERENCES

Aminu, E.F., (2015). A mechanism for detecting dead URLs in XTM-based ontology repository. *International Journal of Computer Applications*, *11*(12). 6.

Bar-IIan, J. (2005). Comparing rankings of search results on the Web. *Information Processing & Management.* 41(6). 1151-1519. Available: http://www.sciencedirect.com/science/article/pii/S0306457305000312

Baye, M. R., De los Santos, B. and Wildenbeest, M. R. (2016), Search Engine Optimization: What Drives Organic Traffic to Retail Sites? *Journal of Economics & Management* Strategy, 25: 6–31. doi: 10.1111/jems.12141. Availble: http://onlinelibrary.wiley.com/doi/10.1111/jems.12141/full

Besek, J.M. (2003). Copyright issues relevant to the creation of a digital archive: A preliminary assessment. Councile on Library and Information Resources, Washington D.C. and Library of Congress. p. 11 Available: http://files.eric.ed.gov/fulltext/ED480764.pdf

Bilal, D. (2012). Ranking, relevance judgement, and precision of information retrieval on children's queries: Evaluation of Google, Yahoo!, Bing, Yahoo!Kids, and ask Kids. *Journal of the American Society for Information Science and Technology*, *63*(9). 1879.

Choo, C.W. (1993). Environmental scanning: acquisition and use of information by managers. In ME Williams (Ed.). *Annual review of information science and technology*, *28.*

Ericsson, K..A., Krampe, R., Tesch-romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*. 100(3), 363-406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Eveland, W.P. Jr., Marton, M., & Seo, M., (2004). Moving beyond "Just the Facts" The influence of online news on the content and structure of public affairs knowledge. *Communication Research*, *31*(1), 82.

Green, S. (2010). Are website copyright violations hurting newspaper's bottom line? *Las Vegas Sun*. Available: http://lasvegassun.com/news/2010/aug/04/are-website-copyright-violations-hurting-newspaper/

Himelboim, I. (2010). The international network structure of news media: an analysis of hyperlinks usage in news web sites. *Journal of Broadcasting & Electronic Media*, *54*(3). 373-390. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2010.499050

Isbell, K. (2010). The rise of the news aggregator: Legal implications and best practices (August 30, 2010). Berkman Center Research Publication No. 2010-10. Available at SSRN: http://ssrn.com/abstract=1670339 or http://dx.doi.org/10.2139/ssrn.1670339

Lau, R.Y.K., Liao, S.S.Y., Wong, K.F., Chiu, D.K.W. (2012). Web 2.0 envirnmental scanning and adaptive decision support for business mergers and acquisitions. *MIS Quarterly*, *36*(4). 1239-1268.

Liu, S., Turban, E., Lee, M.K.O. (2000). Software agents for environmental scanning in electronic commerce. *Information Systems Frontiers*, *2*(1). 85-98.

Lopez-Taruella, A. (2012). Google and the law. *Empirical approaches to legal aspects of knowledge-economy business models.* T.M.C. Asser Press, The Hague, The Netherlands. p. 117

Morishima, A., Nakamizo, A., Iida, T., Sugimoto, S., & Kitagawa, H. (2009) Bringing your dead links back to life: a comprehensive approach and lessons learned. *Proceedings of the 20th ACM conference on Hyptetext and Hypermedia.* ACM, New York. 15-24.

Ott, S. (2008). Belgian court: Google News violates copyright law. *Links & Law*. Available: http://www.linksandlaw.com/news-update48-google-news-copyright.htm

Perry, D.C., Taylor, M., Doerfel, M.L. (2003). Internet-based communication in crisis management. *Management Communication Quarterly*, *17*(2) 206-232.

Peterson G., Aslani P., Williams K.A., (2003). How do consumers search for and appraise information on medicines on the internet? A qualitative study using focus groups. *Journal of Medical Internet Research*. Available: http://www.jmir.org/2003/4/e33/

Reinhold, D., Wagner, R., Scholz, S.W. (2005). An internet-based approach to environmental scanning in marketing planning. *Marketing Intelligence & Planning*, *23*(2). 189 – 199.

Rennie, S.C. (2016). Providing information on environmental change: Data management, discovery and access in the UK Environmental Change Network Data Center. *Ecological Indicators*. Available: http://www.sciencedirect.com/science/article/pii/S1470160X16300140

Shen, W., Wang, J., Luo, P., Wang, M. (2013). Linking name entities in Tweets with knowledge base via user interest modeling. Proceedings of t 19th ACM SIGKDD. *International Conference on Knowledge Discovery and Data Mining.* ACM, New York. 68-76.

Spink, A., Jansen, B.J., Blakely, C., Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. *Information Processing & Management*, *42*(5). 1379.

Sterling, G. (2014). Strict new copyright law forces end of Google news in Spain. *Search Engine Land*. December 10, 2015. Availble: http://searchengineland.com/responding-strict-new-copyright-law-google-shutter-news-site-spain-210648

Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009). Discovering and Maintaining Links on the Web of Data. *Proceedings of the 8th International Semantic Web Conference*, ISWC 2009, Chantilly, VA, USA, October 25-29, 650-665.

Wang, L., Wang, J., Wang, M., Liang, Y., Xu, D. (2012). Using internet search engines to obtain medical information: A comparative study. *Journal of Medical Internet Research*, *14*(3). Available: http://www.jmir.org/2012/3/e74/

Webb, G.K.. (2016). Public management decisions related to the decline of California deer populations: A comparative management approach. *Environment and Ecology Research*, *4*(2). 63-73.

Zaharoff, H.G. (2001). A writer's guide to fair use. *Writer's Digest*. January. Available: http://www.mbbp.com/news/writers-guide-to-fair-use