

Similarity Among Web Pages Based on their Link Structure

Jesús Ubaldo Quevedo and S. H. Stephen Huang
Department of Computer Science
University of Houston
Houston, TX 77204-3010

e-mail: jquevedo@cs.uh.edu

ABSTRACT

Handwritten: topic "t"

Hyperlinks among web pages enclose indeed a great deal of human assessment [14]. For instance, if page p contains a link to page q , then the person designing page p is recommending for some reason the visitation of page q . Perhaps both pages share a common interest in a particular topic, and p is judging ~~a~~ valuable the information contained in q . Moreover, p may be recommending more than one page, in that case, are those pages recommended by p alike? What about if more persons interested in t suggest also those same pages recommended by p ? Our notion of similarity will take advantage of these human findings by measuring them.

KEY WORDS

Handwritten: on this topic "t"

Related pages, similarity, web retrieval, search, WWW, ranking

1. INTRODUCTION

Nowadays, the Internet is accessible to almost everyone worldwide to post and retrieve any kind of document. There are different tools that make easy to anyone to post whatever he or she wants on the net. Therefore, the Internet has become the world largest collection of information and data, which makes very difficult to find precise and valuable information. The traditional way of exploring the web is by a search engine such as [1], [8], [17] and [25]. When prompting them with a specific query, it is very often the case of getting millions of documents as an answer, which goes beyond human capabilities. Search engines [1], [8], [17], [25] implement all kind of known Information Retrieval (IR) techniques [2], [9], [13], [15], [24] to access

data from the web. However, it has been seen during the past years that those methods were not enough to get valuable information from the WWW [4], [11], [12], [14], [21], [23], [26], [3], [7], [10].

There have been several attempts to improve the quality of the information retrieval from the web [4], [12], [14], [15], [16], [18], [19], [21], [22], [23], [26], [3], [6], [7], [10]. These approaches have considered other aspects of the objects found in the WWW such as link structure and tag analysis. In these efforts for manipulating and extracting information from the web, we could distinguish various types of problems addressed like web query languages [4], [16], [21], [22], relevance of a given document with respect to the query [12], [14], [18], [23], [26], [3], [5], [19] and similarity or related pages [14], [23], [5], [20], [7], [10].

Our paper focus on the similarity area, addressing the following queries:

- How similar is page p_1 with respect to page p_2 ?
- How similar is page p_1 with respect to the group of pages P ?
- How similar are the pages in the set P among themselves?

We attempt to answer the previous inquiries by giving a number that represents the "measure of similarity".

Applications. Let us now illustrate the usefulness of the effective solution of those queries to solve practical problems while searching for information on the web.

Example 1: A user gets millions of url's from a search engine for a given query. After examining some url's, the user finds a document (or documents) that he considers solves his petition appropriately. Now, the user can employ our algorithm to go to other pages and get those "similar" to the one(s) chosen initially.

Example 2: After sometime, a unique user has acquired a collection of url's related to a specific topic. Now, he wants to find more like the one he already have. The user can utilize our algorithm to measure the similarity of the group of pages he has already in his possession. Then, he would use it as a standard that a page needs to meet in order to be part of the initial group.

Notice that in these examples, the similarity-search is based on the personal preference of the user set by the initial page or group of pages chosen. The user can control the "similarity-measure" or "closeness" of the documents desired. Moreover, the algorithm goes through all the pages provided by the search engine pruning all unrelated documents accordingly with the user's preferences. Therefore, the user does not need to go manually to the sometimes-impossible task of reading all documents.

Our algorithm is based on the link structure of the documents on the WWW. Our discussion is based on the fact that web designers have already conferred some recognition to other pages by placing a link in their document [14]. Moreover, we say that the designer of the web page is judging somehow similar two documents when they treat the same topic as the page he is designing, and he is placing a link to them. Hence, we will take advantage of the human judgment to satisfy human judgment.

2. RELATED WORK

Kleinberg [14] in 1998 introduced the concepts of "authorities" and "hubs". Perhaps his paper is one of the most widely cited in the areas of hyperlinked environments. His article presented an analysis of the link structure that states that a link recognizes authority of the other document. The main statement

is that those conferring the recognition are called "hubs" and those receiving the recognition are called "authorities". He did utilize his algorithm to solve the "similar-page queries", but he did not present an actual way of getting a measure as we do here.

There are other systems that find "Similar Pages", online, like Google with its "similar Pages" feature [8], Netscape with its "what's related?" option [20], and TOPIC from the University of Toronto [19], [18]. As well as some research in finding and evaluating "related pages" or "similarity search on the web" such as [7] and [10]. Nevertheless, none of them present the concepts and techniques used here.

3. SIMILARITY MEASURE

NOTATION:

A link relationship between pages p and q is written as $p \rightarrow q$, which can be seen as:

- There is a direct link from page p to page q or $\text{link}(p,q,1)$.
- Page p recommends page q.
- Page p is one reference of page q.
- Page p points to page q.
- Page q is reachable from page p.

A link relationship $p \rightarrow p$, is interpreted as:

- The empty link or $\text{link}(p,p,0)$.
- Page p is self-recommending.

A path expression $p_1 \rightarrow p_2 \rightarrow p_3$, is interpreted as:

- There is an indirect link from page p1 to page p3 or $\text{link}(p_1,p_3,2)$.
- There are two direct links in this path expression, $\text{link}(p_1,p_2,1)$ and $\text{link}(p_2,p_3,1)$.
- Page p3 is transitively recommended by page p1.
- Page p3 is transitively referenced by page p1.
- Page p3 is reachable from page p1.

A general path expression $p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$, is represented as follows:

- $p_1 \Rightarrow p_n$ or $\text{link}(p_1, p_n, n-1)$, where $n-1$ is positive integer

Definition: References of page q.

The set $R(q)$ will be all pages p for which $p \rightarrow q$ is true.

The set $R(q)$ is decomposed, using the idea of Prestige [Ref], into three sets $R_1(q)$, $R_2(q)$, and $R_3(q)$. $R_1(q)$ are those pages related to the topic t treated by q , while $R_2(q)$ are pages not related to topic t , and $R_3(q)$ are pages enclosed in the same domain as page q is located. $R_1(q)$ and $R_2(q)$ must be located at a different server from $R_3(q)$. Therefore, $R(q) = R_1(q) \cup R_2(q) \cup R_3(q)$, and $R_1(q) \cap R_2(q) \cap R_3(q) \neq \emptyset$.

Definition: Closure of page p

The set $C^+(p)$ will be all pages reachable from p or all pages q for which $p \Rightarrow q$ is true.

The closure of a set of pages $S = \{p_1, \dots, p_n\}$ is defined as $C^+(S) = C^+(p_1) \cup \dots \cup C^+(p_n)$.

Definition: Closed set of pages

Let S be a set of pages. S is a closed set of pages iff $C^+(S) = S$.

Definition: Source

A group of pages K form a Source for the set of pages contained in S , when all pages in S are reachable from K .

Definition: Minimal Source

The set CK form a Minimal Source for S when the removal of any pages from CK , makes CK no longer a Source.

Definition: Prestige of page p

Prestige is the measure of approval given to a particular document by other creators of web pages [21], [22]. It is totally based on the references of p , $R(p)$.

$$\text{Prestige}(p) = W_1 * \|R_1(p)\| + W_2 * \|R_2(p)\| + W_3 * \|R_3(p)\|$$

W_1 , W_2 and W_3 weight the level of importance of the references. R_1 are pages from different sever and dealing with the same topic t , these are recommendations from documents in the same field; therefore, W_1 should have the highest weight. R_2

are pages from different server and not treating the topic t , these are recommendations of documents not experts in t . W_2 should have a weight less than W_1 . While R_3 are pages in the same domain, these are self-recommending pages. We want to avoid self-recommendations that may lead us to Web Spamming. W_3 should have a small negative value that penalizes self-recommendations, which should not affect honest self-references. However, a Web spammer adding several links points to itself should not be classified as prestigious page. Suggested values for these weights are $W_1=1$, $W_2=0.1$ and $W_3=-0.01$. Notice that W_3 would not really affect the Prestige computation unless $\|R_3(p)\|$ is large.

Definition: Plain Similarity

Plain similarity is based on the percentage of references that would have in common two pages.

$$\text{Sim}(p_1, p_2) = \frac{\|R_1(p_1) \cap R_1(p_2)\|}{\|R_1(p_1) \cup R_1(p_2)\|}$$

Definition: Similarity with Prestige

Similarity with Prestige considers all types of recommendations, and applies the corresponding weights.

$$\text{Sim}_1(p_1, p_2) = \frac{\|R_1(p_1) \cap R_1(p_2)\|}{\|R_1(p_1) \cup R_1(p_2)\|}$$

$$\text{Sim}_2(p_1, p_2) = \frac{\|R_2(p_1) \cap R_2(p_2)\|}{\|R_2(p_1) \cup R_2(p_2)\|}$$

$$\text{Sim}_3(p_1, p_2) = \frac{\|R_3(p_1) \cap R_3(p_2)\|}{\|R_3(p_1) \cup R_3(p_2)\|}$$

$$\text{Sim}_P(p_1, p_2) = W_1 * \text{Sim}_1(p_1, p_2) + W_2 * \text{Sim}_2(p_1, p_2) + W_3 * \text{Sim}_3(p_1, p_2)$$

Definition : Confidence

Consider that pages p_1 and p_2 are related to topic t , and S is the set of all pages dealing with t . The confidence of $\text{Sim}(p_1, p_2)$ written as $\text{Conf}(\text{Sim}(p_1, p_2))$ will be 100% if $K = R_1(p_1) \cup R_1(p_2)$ is a Source of S ; otherwise, $\text{Conf}(\text{Sim}(p_1, p_2)) = \frac{\|K \cap CK\|}{\|CK\|}$ where CK is a Minimal Source for S . K is a Source of S if $S = C^+(p_1) \cup C^+(p_2)$.

Definition: Inverse Similarity

Inverse similarity uses the closure function to determine the percentage of common recommendations.

$$\text{InvSim}(p1,p2) = \frac{\|C+(p1) \cap C+(p2)\|}{\|C+(p1) \cup C+(p2)\|}$$

All functions defining similarity can be extended to compute the similarity among n pages, for instance, plain similarity among n documents can be obtained from the evaluation of the following expression:

$$\text{Sim}(p1, \dots, pn) = \frac{\|R1(p1) \cap \dots \cap R1(pn)\|}{\|R1(p1) \cup \dots \cup R1(pn)\|}$$

This expression tells how similar those documents among themselves are. It is needed an additional expression to compare the similarity of a particular page p with respect to a selected group of pages $S=\{p1,p2,\dots,pn\}$ such as $\text{SimTo}(p/p1,p2,\dots,p3)$.

$$\text{SimTo}(p/p1) = \frac{\|R1(p) \cap R1(p1)\|}{\|R1(p1)\|}$$

$\text{SimTo}(p/p1)$ computes how similar is page p with respect to page p1, while $\text{SimTo}(p1/p)$ computes how similar is p1 with respect to p.

$$\begin{aligned} \text{SimTo}(p1/p2,p3,\dots,pn) &= A / B \\ \text{Where:} \\ A &= \|R1(p1) \cap R1(p2) \cap R1(p3) \\ &\quad \cap \dots \cap R1(pn)\|, \text{ and} \\ B &= \|R1(p2) \cap R1(p3) \cap \dots \cap \\ &\quad R1(pn)\| \end{aligned}$$

$\text{SimTo}(p1/p2,p3,\dots,pn)$ computes how similar is p1 with respect to the group of pages defined by $\{p2,p3,\dots,pn\}$.

4. EXPERIMENTS

The examples shown on this section are based on the information provided by Table 1.

Example 1:

P1: www.rational.com
P3: www.objectsbydesign.com

$$\begin{aligned} \|R1(p1) \cap R1(p3)\| &= 5 \\ \|R1(p1) \cup R1(p3)\| &= 37 \\ \text{Sim}(p1,p3) &= \frac{\|R1(p1) \cap R1(p3)\|}{\|R1(p1) \cup R1(p3)\|} = 5/37 \approx 14\% \\ \text{SimTo}(p1/p3) &= \frac{\|R1(p1) \cap R1(p3)\|}{\|R1(p3)\|} = 5/9 \approx 56\% \\ \text{SimTo}(p3/p1) &= \frac{\|R1(p1) \cap R1(p3)\|}{\|R1(p1)\|} = 5/33 \approx 15.2\% \end{aligned}$$

-> **Elements of $R1(p1) \cap R1(p3)$:**

- jodi.ecs.soton.ac.uk
- plg.uwaterloo.ca
- www.cetus-links.org
- www.dsic.upv.es
- www.omg.org

Example 2:

P1: www.rational.com
P4: www.cs.york.ac.uk/uml2000

$$\begin{aligned} \|R1(p1) \cap R1(p4)\| &= 6 \\ \|R1(p1) \cup R1(p4)\| &= 33 \\ \text{Sim}(p1,p4) &= \frac{\|R1(p1) \cap R1(p4)\|}{\|R1(p1) \cup R1(p4)\|} = 6/33 \approx 18.2\% \\ \text{SimTo}(p1/p4) &= \frac{\|R1(p1) \cap R1(p4)\|}{\|R1(p4)\|} = 6/6 = 100\% \\ \text{SimTo}(p4/p1) &= \frac{\|R1(p1) \cap R1(p4)\|}{\|R1(p1)\|} = 6/33 \approx 18.2\% \end{aligned}$$

-> **Elements of $R1(p1) \cap R1(p4)$:**

- jodi.ecs.soton.ac.uk
- www.cetus-links.org
- www.db.informatik.uni-bremen.de
- www.dcs.ed.ac.uk
- www.jeckle.de
- www.onesmartclick.com

Example 3:

P1: www.rational.com
P2: www.omg.org

$$\begin{aligned} \|R1(p1) \cap R1(p2)\| &= 18 \\ \|R1(p1) \cup R1(p2)\| &= 42 \\ \text{Sim}(p1,p2) &= \frac{\|R1(p1) \cap R1(p2)\|}{\|R1(p1) \cup R1(p2)\|} = 18/42 \approx 43\% \end{aligned}$$

$$\text{SimTo}(p1/p2) = || R1(p1) \cap R1(p2) || / ||$$

$$R1(p2) || = 18/27 \approx 67\%$$

$$\text{SimTo}(p2/p1) = || R1(p1) \cap R1(p2) || / ||$$

$$R1(p1) || = 18/33 \approx 55\%$$

<p>Page p1= www.rational.com</p> <p>-> Elements of R1(p1):</p> <ul style="list-style-type: none"> - ivs.cs.uni-magdeburg.de - jodi.ecs.soton.ac.uk - plg.uwaterloo.ca - usuarios.dialdata.com.br - www.agilemodeling.com - www.ajug.org - www.ambysoft.com - www.analisi-disegno.com - www.cc.ioc.ee - www.celigent.com - www.cetus-links.org - www.conallen.com - www.cs.toronto.edu - www.cs.ut.ee - www.cs.york.ac.uk - www.csci.csusb.edu - www.db.informatik.uni-bremen.de - www.dcs.ed.ac.uk - www.dsic.upv.es - www.jaist.ac.jp - www.jeckle.de - www.jeffsutherland.org - www.jguru.com - www.michael-thomas.com - www.objekttechnik.se - www.omg.org - www.onesmartclick.com - www.rational.co.jp - www.smartdraw.com - www.stm.tj - www.therationaledge.com - www.uml.crespim.uha.fr - www.univ-pau.fr <p> R1(p1) = 33</p>	<p>> Page p2 = www.omg.org</p> <p>-> Elements of R1(p2):</p> <ul style="list-style-type: none"> - ivs.cs.uni-magdeburg.de - jodi.ecs.soton.ac.uk - usuarios.dialdata.com.br - www-db.stanford.edu - www.ambysoft.com - www.celigent.com - www.cetus-links.org - www.cs.toronto.edu - www.cs.ut.ee - www.cs.york.ac.uk - www.db.informatik.uni-bremen.de - www.dcs.ed.ac.uk - www.dsic.upv.es - www.embarcadero.com - www.jaist.ac.jp - www.jeckle.de - www.jguru.com - www.michael-thomas.com - www.modelingstyle.info - www.nofusion.com - www.onesmartclick.com - www.rational.com - www.smartdraw.com - www.sparxsystems.com.au - www.uml.org - www.univ-pau.fr - www.visualobject.com <p> R1(p2) = 27</p>
<p>> Page p3= www.objectsbydesign.com</p> <p>-> Elements of R1(p3):</p> <ul style="list-style-type: none"> - jodi.ecs.soton.ac.uk - nsuml.sourceforge.net - opensource.objectsbydesign.com - plg.uwaterloo.ca - www.cetus-links.org - www.dsic.upv.es - www.embarcadero.com - www.omg.org - www.onesmartclick.com <p> R1(p3) = 9</p>	<p>> Page p4= www.cs.york.ac.uk</p> <p>-> Elements of R1(p4):</p> <ul style="list-style-type: none"> - jodi.ecs.soton.ac.uk - www.cetus-links.org - www.db.informatik.uni-bremen.de - www.dcs.ed.ac.uk - www.jeckle.de - www.onesmartclick.com <p> R1(p4) = 6</p>

Table 1: Four pages related to UML with their R1 elements

-> Elements of $R1(p1) \cap R1(p2)$:

- ivs.cs.uni-magdeburg.de

- jodi.ecs.soton.ac.uk
- usuarios.dialdata.com.br
- www.ambysoft.com

- www.celigent.com
- www.cetus-links.org
- www.cs.toronto.edu

- www.cs.ut.ee
- www.cs.york.ac.uk
- www.db.informatik.uni-bremen.de
- www.dcs.ed.ac.uk
- www.dsic.upv.es
- www.jeckle.de
- www.jguru.com
- www.michael-thomas.com
- www.onesmartclick.com
- www.smartdraw.com
- www.univ-pau.fr

5. CONCLUSIONS AND FUTURE WORK

This work has presented an innovative technique for computing similarity among web pages. It has the capability of comparing a page with another page or comparing a page with group of pages or the similarity among the pages contained in one particular set. These results can be used to rank pages as well as finding related pages. We have compared our results with those obtained from "related pages" from Google, it seems we could get an improvement of about 40% of accuracy. Those results will be shortly available through our website at <http://www.cs.uh.edu/~jquevedo/similarity/>.

So far our algorithm uses only the link structure to compute similarity, but other aspects of the web may be incorporated such as "tag analysis" and textual information to compute a composite score in similarity.

6. ACKNOWLEDGEMENT

Our gratitude to: José-Luis Torres-Pérez and Carlos-Alberto Torres-Pérez (graduate students at Universidad Autónoma de Guadalajara) for their experiments and verification of our results.

REFERENCES

- [1] AltaVista. AltaVista Search Engine, <http://www.altavista.com>.
- [2] Baeza-Yates, R., and Ribeiro-Neto, B. Modern Information Retrieval, Addison-Wesley, 1999.
- [3] Bharat, K., and Henzinger, M. Improved algorithms for topic distillation in hyperlinked environments. *Proceedings of the 21st International ACM SIGIR Conference on Research Development in Information Retrieval (SIGIR '98)*, pp. 104-111, 1998.
- [4] Carriere, J., and Kazman, R. WebQuery: Searching and visualizing the Web through connectivity, Proceedings of the Sixth International Conference on the World Wide Web, Santa Clara CA., 1997.
- [5] Chakrabarti, S., Dom, B., Gibson, D., and Indyk, P. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference of Management of Data*, pp. 307-318, 1998.
- [6] Chakrabarti, S., Dom, B., Gibson, D., Kumar, S., R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Experiments in topic distillation. In *ACM-SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*, 1998.
- [7] Dean, J., Hensing, M. Finding related pages in the world wide web. In *Proceedings of WWW8*, 1999.
- [8] Google. Google Search Engine at Google.com. <http://www.google.com>.
- [9] Grossman, D.A., and Frieder O., Information Retrieval: Algorithms and Heuristics. Kluwer, August 1998.
- [10] Haveliwala, T., H., Glonis, A., Klein, D., Indyk, P. Evaluating strategies for similarity on the web. In *Proceedings of WWW11*, 2002.
- [11] Huberman, B. A. and Adamic, L. A. Evolutionary Dynamics of the World Wide Web. Xerox Palo Alto Research Center. February 1999.
- [12] Keast, G., Toms, E. G., and Cherry, J. Measuring the reputation of web sites. Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, January 2001.
- [13] Kleinberg, J., and Tomkins, A. Applications of linear algebra in information retrieval an hypertext. Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, May 1999.
- [14] Kleinberg, J., M., Authoritative Sources in a Hyperlinked Environment, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, 1998.
- [15] Kobayashi, M., and Takeda, K. Information Retrieval on the Web ACM Computing Surveys, 2000.

- [16] Lakshmanan, L., Sadri, F., and Subramanian, I., A Declarative Language for Querying and Restructuring the Web. In Proc. 6th Int. workshop on research Issues in Data Engineering, New Orleans, 1996.
- [17] Lycos. Lycos Web Site, <http://www.lycos.com>.
- [18] Mendelzon, A. O., and Rafiei, D. What do the Neighbors Think? Computing Web Page Reputations. IEEE Data Engineering Bulletin, September 2000.
- [19] Mendelzon, A., O., and Rafiei, D., An Autonomous Page Ranking Method for Metasearch Engines, Proceedings of the 11th International World Wide Web conference, Honolulu, Hawaii, May 7-11 2002.
- [20] Netscape Communications Corporation. 'What's Related FAQ' web page. <http://home.netscape.com/escapes/related/faq.html>
- [21] Quevedo, J., U., Covarrubias, A., G., and Huang, S., Improving Retrieval by Querying and Examining Prestige. Proceedings of the 11th International World Wide Web conference, Honolulu, Hawaii, May 7-11 2002.
- [22] Quevedo, J., U., Covarrubias, A., G., and Huang, S., Querying and Ranking Web Documents with Prestige. Proceedings of the International Conference on Information and Knowledge Engineering (IKE'02), Las Vegas, Nevada, June 24-27 2002, pp. 443-446.
- [23] Quevedo, J., U., Huang, S., and Vargas, M., L., TAKER: Improving Retrieval by Exploring Tag-Keyword Relationships. Proceedings of the International Conference on Information and Knowledge Engineering (IKE'02), Las Vegas, Nevada, June 24-27 2002, pp. 476-481.
- [24] Salton, G., and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
- [25] Yahoo. Yahoo search engine at Yahoo.com, <http://www.yahoo.com>.
- [26] Zhang, D., and Dong, Y. An Efficient Algorithm to Rank Web Resources. Proceedings of the 9th international World Wide Web conference on Computer networks: The international journal of computer and telecommunications networking, June 2000, pp. 449-455.