A Hyperlink-Based Recommender System Written in Squeal

Ellen Spertus
Mills College
5000 MacArthur Blvd.
Oakland, CA 94613 USA
spertus@mills.edu

Lynn Andrea Stein
MIT AI Lab
545 Technology Square
Cambridge, MA 02139 USA
las@ai.mit.edu

1. ABSTRACT

Human beings, not machines, are the ultimate experts for information retrieval tasks, including recommender systems. Consequently, computers are most useful when they combine information about people's judgments. Collaborative filtering systems make use of this observation by having users explicitly rate items, such as Web pages, with the system making recommendations to other users based on overlapping areas of interest. A disadvantage of collaborative filtering, at least as currently implemented, is that it depends on users' explicitly entering data, which can be inconvenient and time-consuming. We describe the design, implementation, and performance of a recommender system that works by mining publicly-available hyperlinks on the Web, producing results competitive with the best text-based system. We also demonstrate the utility of the Squeal language for structure-based Web queries.

1.1 Keywords

Recommender system, semi-structured information, Web, relational databases, SQL, Squeal

2. INTRODUCTION

"We will now take up one of the most interesting branches of the calculus of the passions: the art of enabling anyone anywhere, even in places where he is a total stranger, to make instant contact with people with whom he is in complete sympathy. If the theory of attraction offered no other advantage, would it not still be a boon to all of mankind?

— Charles Fourier (1858) [1 , p. 378]

One useful class of information retrieval applications is recommender systems [3], where a program recommends new Web pages (or some other resource) judged likely to be of interest to a user, based on the user's initial set of seed pages P. A standard technique for recommender systems, used by the Excite search service (www.excite.com), is extracting keywords that appear on the seed pages and returning pages that contain these keywords. Note that this technique is based purely on the text of a page, independent of any inter- or intra-document structure.

Another technique for making recommendations is collaborative filtering [6], where pages are recommended that were liked by other people who liked P. This is based on the observation that items thought valuable/similar by one user are likely to by another user. As collaborative filtering is currently practiced, users explicitly rate pages to indicate their recommendations. This inconvenient and expensive step can be eliminated through data mining by interpreting the act of creating hyperlinks to a page as being an implicit recommendation. In other words, if a person links to pages Q and R, we can guess that people who like Q may like R, especially if the links to Q and R appear near each other on the referencing page (such as within the same list). This approach takes advantage of inter-document structure (i.e., hyperlinks) and intra-document structures (e.g., lists). We call our application a *ParaSite* because it makes use of information on Web pages in ways unintended by their authors.

Accordingly, if a user requests a page similar to a set of pages {P1, …, Pn}, the system can find pages R that point to a maximal subset of these pages[1] and then return to the user what other pages are referenced by R. Note that our ParaSite does not have to understand what the pages have in common. We assume that co-reference implies some sort of human-defined similarity.

For example, suppose a user wishes to find pages similar to the Association for Computing Machinery site (www.acm.org) and the Association for Artificial Intelligence site (www.aaai.org). A page at MIT entitled "Computer Science Web Sites" (http://libraries.mit.edu/

[1] The AltaVista search service (www.altavista.digital.com) supports a keyword "link" that can be used to request pages that link to a specified page.

barker/Subjects/CS/ComputerSciWeb.html), excerpted in Figure 1, points to both of these sites. It also points to the Computing Research Association (www.cra.org), which is indeed related. A way to avoid less-relevant links, such as those appearing under the heading "Fun Pages", is taking intra-document structure into account, as discussed later in the paper.

Note the similarity to *bibliometrics*, the statistical study of documents, which includes citation indexing [5]. *Co-citation* refers to when two papers are referenced by a common source [7] and is equivalent to the ParaSite's judging two web documents similar if they are both pointed to by the same page. Brewster Kahle uses the term "siblink" to refer to such pages. The term "sitation" has been coined by Gerry McKiernan to describe the study of links among Web pages [4]. Jon Kleinberg and his colleagues have developed algorithms to find high-quality Web pages by examining link hierarchy [2].

3. BACKGROUND

Our recommender system is written in Squeal [8], a system we developed for making queries on the Web in Structured Query Language (SQL). The Squeal relations used for the recommender system appear in Figure 2. We represent relation names in SMALL CAPS, column names in **bold face**, and parameters in *italics*. The VALSTRING relation is used to associate an integer, **value_id**, with a string, **textvalue**; the integer is used in other tables to refer to the string. The URLS relation is used to map one or more **value_id**s representing URL strings to a unique **url_id**. It is necessary because the same page can be referred to by multiple URL strings (e.g., "www.ai.mit.edu" and "www.ai.mit.edu/index.html"). The LINK relation is used to represent hyperlinks. The **source_url_id**, **anchor_value_id**, and **dest_url_id** indicate the page on which a link occurs, its anchor text, and its destination.

For example, the following query asks what pages are pointed to by "www.ai.mit.edu":

select *vdest*.**textvalue**
from VALSTRING *vsource*, VALSTRING *vdest*, URL *usource*, URL *udest*, LINK *l*
where *vsource*.**textvalue** = "www.ai.mit.edu"
and *usource*.**value_id** = *vsource*.**value_id**
and *l*.**source_url_id** = *usource*.**url_id**
and *udest*.**url_id** = *l*.**dest_url_id**
and *vdest*.**value_id** = *udest*.**value_id**

How Squeal determines this information is described elsewhere [8].

The **hstruct** and **lstruct** relations indicate where in the page's header and list hierarchy the link appears; for example, under the first H1 header and in a doubly-nested list. Figure 3 shows the portions of the LINK, URLS, and VALSTRING tables for the information from the page shown in Figure 1. Note the different **hstruct** and **lstruct** values for links appearing beneath different headers and on different lists.

4. IMPLEMENTATION

We use the following algorithm to find pages similar to P1 and P2:

1. Generate a list of pages R that point to P1 and P2.
2. List the pages most commonly pointed to by pages within R.

The Squeal code for the algorithm is shown in Figure 4. Some heuristics for improving precision are:

1. Only return target pages that include a keyword specified by the user.
2. Only return target pages that point to one or both of P1 and P2.
3. Only follow links that appear in the same list and under the same header as the links to P1 and P2.

This last heuristic was motivated by the observation that

Computer Science Web Sites

Organizations and Institutions in Computer Science

ACM - Association for Computing Machinery
AAAI - American Association for Artificial Intelligence
CRA - Computing Research Organization

Fun Sites

The Ada Project
Lost Library of MOO

Figure 1: An excerpt from "http://libraries.mit.edu/barker/Subjects/CS/ComputerSciWeb.html"

VALSTRING

colname	*type*
value_id	int
textvalue	text

URLS

colname	*type*
url_id	int
value_id	int
variant	int

LINK

colname	*type*
source_url_id	int
anchor_value	int
dest_url_id	int
hstruct	binary
lstruct	binary

Figure 2: Definitions of the VALSTRING, URLS, and LINK relations

LINK

source_url_id	anchor_value_id	dest_url_id	hstruct	lstruct
1	201	2	1 1 0 0 0 0	1 0 0 0 0 0
1	202	3	1 1 0 0 0 0	1 0 0 0 0 0
1	203	4	1 1 0 0 0 0	1 0 0 0 0 0
1	204	5	1 2 0 0 0 0	2 0 0 0 0 0

VALSTRING

value_id	textvalue
201	ACM - Association for Computing Machinery
202	AIAA - American Association for Artificial Intelligence
203	CRA - Computing Research Organization
204	The Ada Project
…	
301	libraries.mit.edu/barker/Subjects/CS/ComputerSciWeb.html
302	www.acm.org
303	www.aaai.org
304	www.cra.org
305	www.cs.yale.edu/HTML/YALE/CS/HyPlans/tap/tap.html

URLS

url_id	value_id
1	301
2	302
3	303
4	304
5	305

Figure 3: Portions of the LINK, URLS, and VALSTRING relations

some pages contains hundreds or thousands of links and that the most similar pairs of links are likely to be within the same list or under the same header. In our example, this would exclude the links to "The Ada Project" and "Lost Library of Moo", which indeed are less relevant. All of these heuristics can be expressed easily in Squeal [8].

5. EVALUATION

5.1 Method

To compare the structure-based and text-based approaches, we used the above ParaSite with the last heuristic and the Excite "more like this" feature. Because Excite can only find pages similar to a single page, not a set of them, we only provided a single URL to each system for each round of the test.

We had four human subjects submit a set of seed URLs that interested them. For thirteen URLs given, we provided users with the top 5 recommendations of each system, which users then rated on a scale of 0 to 4 for relevance, interestingness, and novelty. As subjects pointed out, a rating for novelty seems not to be applicable when a page was entirely irrelevant. For this reason, when "averaging" ratings, novelty was treated as zero when relevance was zero. Full details about the experiment appear elsewhere [8].

5.2 Results

On average, the Excite pages were judged more relevant (1.84 vs. 1.36) and interesting (1.63 vs. 1.47) than the ParaSite pages, while the ParaSite pages were judged more novel (1.32 vs. 1.12). The results of the evaluation of each set of recommendations can be divided into three cases: those where all the ParaSite averages were higher (3), where the Excite averages were higher (4), and where the results are mixed (6). Here, we discuss one of the cases with mixed results.

5.3. Example

The URLs returned by each system for the "Geek Site of the Day" (www.owlnet.rice.edu/~indigo/gsotd/) are shown in Figure 4. Because ParaSite only made four recommendations, only the top four Excite recommendations are listed. Two of the Excite recommendations were articles about GSotD, one was a review of GSotD and similar sites, and one was a GSotD archive. The ParaSite selections were more diverse: the first two were collections of cool/useless pages, the next was the home page of "CNET: The Computer Network", and the fourth was the Museum of Bad Art. The Excite pages were considered more relevant (1.94 vs. 1.71), while the ParaSite pages were considered more interesting (1.83 vs. 1.44) and novel (2.13 vs. .94). Users disagreed in their written comments as to which system was preferable:

> "System A [Excite] came up with one good suggestion. System B [ParaSite] came up with several. System B [ParaSite] wins…" – P
>
> "I assume the person wants sites that would be interesting or funny to the computer geek, such as things in poor taste. In this case I would choose system A [Excite]." – W

Page	Average r	i	n
Excite recommendations			
WebCrawler review (www.webcrawler.com/News/site06.html)	2.25	1.75	1.25
PC Novice mention (www.owlnet.rice.edu/indigo/gsotd/pcnovice.html)	1.5	.75	0
GSotD, Sep. 1995 (www.owlnet.rice.edu/indigo/gsotd/sept95.htm)	2.25	1.75	.75
News Herald review (www.newsherald.com/BUSINESS/B20.htm)	1.75	1.5	1.75
Excite averages	**1.94**	**1.44**	**.94**
ParaSite Recommendations			
Cool Site of the Day (cool.infi.net)	2	2.25	2
Useless Pages (www.go2net.com/internet/useless/)	2.08	2.08	2.5
CNET: The Computer Network (www.cnet.com)	1.75	1.75	2
Museum of Bad Art (www.glyphs.com/moba/)	1	1.25	2
ParaSite averages	**1.71**	**1.83**	**2.13**

Figure 4: User ratings of recommendations for the Geek Site of the Day (GSotD) seed page. The letters "r", "i", and "n" stand for "relevance", "interestingness", and "novelty".

6. DISCUSSION

The ParaSite suggestions were generally judged more novel, while the Excite ratings were judged more relevant and interesting. There were cases in which each system was markedly superior to the other, and cases in which it was not clear which system was better. Some possible conclusions are:

1. The text-based approach is likelier than the structure-based approach to stay within the seed web site, yielding pages that users find more relevant but less novel.
2. Neither of the two approaches is always superior. Whether the text- or structure-based approach is better depends on the type of link and the user's purpose.
3. A superior system could be built by combining the two approaches.
4. The structure-based approach would have generated more useful results if more pages had been examined for each seed URL.

Further evaluation is planned, with a larger number of users and comparison to collaborative filtering and hybrid techniques.

It is not clear what questions should be asked in evaluating a recommender system and how they interact. For example, novelty in a returned page is only valuable if the page is relevant. The relation between these two metrics and interestingness is less clear. Issues also arise from whether users are rating recommendations of pages selected by themselves or by others. Another evaluation challenge is that users may have task-specific preferences. On some occasions, they might want to find something previously unknown to them, while on others they may be trying to refind a page very similar to their seed page. We are experimenting with options to cover these and other circumstances.

The ease of writing the recommender system suggests that Squeal is useful for this type of application. Other applications that have been built are a personal home page finder and a moved page finder [8]. We are planning to make Squeal publicly available.

We were assisted in this research by Oren Etzioni, Keith Golden, Ken Haase, Tom Knight, and Pattie Maes.

7. REFERENCES

[1] Charles Fourier. *The Utopian Vision of Charles Fourier; selected texts on work, love, and passionate attraction*. Translated, edited, and with an introduction by Jonathan Beecher and Richard Bienvenu Beacon Press, 1971.

[2] Jon Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[3] Paul Resnick and Hal R. Varian. Recommender systems (introduction to special section). *Communications of the ACM*, 40(3):56-58, March 1997.

[4] Ronald Rousseau. Sitations: an exploratory study. *Cybermetrics*, 1(1), 1997. http://www.cin-doc.csic.es/cybermetrics/articles/v1i1p1.html.

[5] Jacques Savoy. Citation schemes in hypertext information retrieval. In Maristella Agosti and Alan F. Smeaton, editors, *Information Retrieval and Hypertex*t, pages 99-120. Kluwer Academic Press, 1996.

[6] Upendra Shardanand and Pattie Maes. Social information filtering: Algorithms for automating "word of mouth". In Computer-Human Interaction (CHI), 1995.

[7] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265-269, 1973.

[8] Ellen Spertus. ParaSite: Mining the Structural Information on the World-Wide Web. PhD Thesis, Department of EECS, MIT, Cambridge, MA, February 1998.

[9] Ellen Spertus and Lynn Andrea Stein. Just-In-Time Databases and the World-Wide Web, Seventh International ACM Conference on Information and Knowledge Management, November 1998.