# Mining the Web's Hyperlinks for Recommendations

Ellen Spertus
Mills College
5000 MacArthur Blvd.
Oakland, CA 94613

spertus@mills.edu

Lynn Andrea Stein
MIT AI Lab
545 Technology Square
Cambridge, MA 02139

las@ai.mit.edu

## 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 a recommender system, which we call ParaSite, that works by mining publicly-available hyperlinks on the Web, producing results competitive with the best text-based system.

## Introduction

One useful class of information retrieval applications is recommender systems [1], 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 [4], 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 mines intra-document structural information.

Accordingly, if a user requests a page similar to a set of pages {*P1, …, Pn*}, the system can find (through AltaVista) pages R that point to a maximal subset of these pages and

then return to the user what other pages are referenced by R. Note that the ParaSite does not have to understand what the pages have in common. It just needs to find a list that includes the pages and can infer that whatever trait they have in common is also exemplified by other pages they point to.

For example, the first page returned by AltaVista that pointed to both Computer Professionals for Social Responsibility ("www.cpsr.org/home.html") and Electronic Privacy Information Center ("www.epic.org") was a list of organizations fighting the Communications Decency Act. This page pointed to the Electronic Frontier Foundation ("www.eff.org") and other related organizations.

Note the similarity to *bibliometrics*, the statistical study of documents, which includes citation indexing [3]. *Co-citation* refers to when two papers are referenced by a common source [5] and is equivalent to the ParaSite's judging two web documents similar if they are both pointed to by the same page. The term "sitation" has been coined by Gerry McKiernan to describe the study of links between Web pages [2].

## 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.

Some heuristics for improving precision are:

1. Only return target pages that include a keyword specified by the user.
2. Return the names of hosts frequently containing referenced.
3. Only return target pages that point to one or both of P1 and P2.
4. 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 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.

## Evaluation

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 provide 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. To capture the simultaneous importance of all three measures, we also computed their products. Full details about the experiment appear elsewhere [7].

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) and had a higher product (4.58 vs. 4.29). 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). We discuss one example in each category.

ParaSite superior: Austin weather

The URLs returned by each system for the page entitled "The Weather Channel – Austin, TX" (www.weather.com/weather/us/cities/TX_Austin.html) are shown in Figure 1. Excite returned Weather Channel reports on other cities in the Southwest, while ParaSite generally returned information, not necessarily weather-related, about Austin or the whole of Texas. Three of the users, including the one who submitted the seed URL, preferred the ParaSite listings. The fourth reviewer thought that the weather information returned by Excite was more relevant. Quantitatively, the ParaSite pages were judged more relevant (1.4 vs. .95), interesting (1.8 vs. .95), and novel (1.37 vs. .25) than the Excite pages. The product of the ratings was much higher for ParaSite than Excite (5.29 vs. .23).

| **P** | | | K | | | W | | | S | | | **Description** | **Avg** | | | **Prod-** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | i | n | r | i | n | r | i | n | r | i | n | | **r** | **i** | **n** | **uct** |
| 0 | 0 | 0 | 1 | 1 | 0 | 3 | 3 | 1 | 0 | 0 | 0 | TWC: Lamesa, TX | 1 | 1 | .25 | .25 |
| 0 | 0 | 0 | 1 | 1 | 0 | 3 | 3 | 1 | 0 | 0 | 0 | TWC: Seminole, TX | 1 | 1 | .25 | .25 |
| 0 | 0 | 0 | 1 | 1 | 0 | 3 | 3 | 1 | 0 | 0 | 0 | TWC: Pecos, TX | 1 | 1 | .25 | .25 |
| 0 | 0 | 0 | 1 | 1 | 0 | 3 | 3 | 1 | 0 | 0 | 0 | TWC: Muleshoe, TX | 1 | 1 | .25 | .25 |
| 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 1 | 0 | 0 | 0 | TWC: Hot Springs, AZ | .75 | .75 | .25 | .14 |
| | | | | | | | | | | | | ***Excite averages*** | **.95** | **.95** | **.25** | **.23** |
| 3 | 3 | 3 | 2 | 3 | 2 | 3 | 3 | 1 | 3 | 2 | 3 | Austin weather | 2.75 | 2.75 | 2.25 | 17.02 |
| 2 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 3 | 2 | 2 | Austin guide | 2 | 2 | 1.75 | 7.00 |
| 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | Texas magazine | 1.25 | 1.75 | 2 | 4.38 |
| 0 | 2 | 2 | 0 | 0 | 3 | 1 | 2 | 1 | 0 | 2 | 0 | Jokes | .25 | 1.5 | .25 | .09 |
| 1 | 2 | 2 | 0 | 0 | 3 | 1 | 2 | 1 | 1 | 1 | 0 | Austin information | .75 | 1.25 | .75 | .70 |
| | | | | | | | | | | | | ***ParaSite averages*** | **1.40** | **1.85** | **1.40** | **5.84** |

Figure 1: User Ratings of Austin Weather Recommendations. "The Weather Channel" is abbreviated "TWC". The four users are indicated by the "P", "K", "W", and "S" columns, and "r", "i", and "n" stand for "relevance", "interestingness", and "novelty", respectively. The user who selected the seed page is shown in boldface.

Excite Superior: MapQuest

The URLs returned by each system for the "MapQuest!" (www.mapquest.com) home page are shown in Figure 2. All 5 sites returned by Excite were highly-relevant map-related sites. The 5 sites returned by ParaSite were all related to travel but much less directly, such as tourist information about San Diego. Excite was rated better for all measures.

| P | | | K | | | W | | | S | | | **Description** | **Avg** | | | **Prod-** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | i | n | r | i | n | r | i | n | r | i | n | | **r** | **i** | **n** | **uct** |
| 2 | 2 | 3 | 2 | 2 | 2 | 3 | 3 | 1 | 3 | 2 | 1 | Geography & Maps | 2.5 | 2.25 | 1.75 | 9.84 |
| 2 | 3 | 2 | 2 | 2 | 2 | 3 | 3 | 1 | 3 | 2 | 2 | AOL NetFind | 2.5 | 2.5 | 1.75 | 10.94 |
| 1 | 2 | 1 | 2 | 2 | 2 | 3 | 2 | 1 | 3 | 2 | 2 | Maps on the Net | 2.25 | 2 | 1.5 | 6.75 |
| 1 | 2 | 3 | 1 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | GeoSystems | 2 | 1.5 | 1.5 | 4.50 |
| 0 | 0 | 0 | 2 | 0 | 0 | 3 | 3 | 1 | 2 | 1 | 0 | MapQuest | 1.75 | 1 | .25 | .44 |
| | | | | | | | | | | | | ***Excite averages*** | **2.20** | **1.85** | **1.35** | **6.49** |
| 0 | 1 | 2 | 1 | 1 | 3 | 3 | 3 | 2 | 3 | 2 | 0 | PCL Map Collection | 1.75 | 1.75 | 1.125 | 3.45 |
| 1 | 1 | 1 | 0 | 0 | 3 | 2 | 3 | 2 | 1 | 3 | 2 | Subway navigator | 1 | 1.75 | 1.25 | 2.19 |
| 0 | 1 | 1 | 0 | 0 | 3 | 3 | 3 | 1 | 2 | 3 | 3 | Xerox Map Viewer | 1.25 | 1.75 | 1 | 2.19 |
| 0 | 1 | 1 | 0 | 0 | 3 | 1 | 2 | 1 | 0 | 1 | 1 | San Diego information | .25 | 1 | .25 | .06 |
| 0 | 1 | 1 | 0 | 0 | 3 | 1 | 2 | 1 | 0 | 0 | 0 | More San Diego info | .25 | .75 | .25 | .05 |
| | | | | | | | | | | | | ***ParaSite averages*** | **.90** | **1.40** | **.78** | **1.59** |

Figure 2: User Ratings of MapQuest Recommendations

Neither System Superior: Geek Site of the Day

The URLs returned by each system for the "Geek Site of the Day" (www.owlnet.rice.edu/~indigo/gsotd/) are shown in Figure 3. 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

| P | | | K | | | W | | | S | | | **Description** | **Avg** | | | **Prod-** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | i | n | r | i | n | r | i | n | r | i | n | | **r** | **i** | **n** | **uct** |
| 1 | 1 | 1 | 2 | 1 | 0 | 3 | 3 | 3 | 3 | 2 | 1 | WebCrawler review | 2.25 | 1.75 | 1.25 | 4.92 |
| 2 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | PC Novice mention | 1.5 | .75 | 0 | 0 |
| 3 | 3 | 0 | 2 | 1 | 0 | 3 | 3 | 3 | 1 | 0 | 0 | GSotD, Sep. 1995 | 2.25 | 1.75 | .75 | 2.95 |
| 3 | 2 | 3 | 1 | 1 | 3 | 0 | 1 | 1 | 3 | 2 | 1 | News Herald review | 1.75 | 1.5 | 1.75 | 4.59 |
| | | | | | | | | | | | | ***Excite averages*** | **1.94** | **1.44** | **.94** | **3.12** |
| 3 | 3 | 3 | 2 | 2 | 2 | 0 | 1 | 2 | 3 | 1 | 2 | Cool Site of the Day | 2 | 2.25 | 2 | 9.00 |
| 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | Useless Pages | 2.08 | 2.08 | 2.5 | 10.85 |
| 3 | 3 | 3 | 0 | 0 | 1 | 1 | 1 | 3 | - | - | - | CNET.COM | 1.75 | 1.75 | 2 | 6.13 |
| 1 | 2 | 3 | 1 | 1 | 3 | 2 | 2 | 2 | 3 | 3 | 2 | Museum of Bad Art | 1 | 1.25 | 2 | 2.50 |
| | | | | | | | | | | | | ***ParaSite averages*** | **1.71** | **1.83** | **2.13** | **7.12** |

Figure 3: User Ratings of Geek Site of the Day (GSotD) Recommendations. Dashes indicate where a user neglected to rate a page.

## Discussion

The ParaSite suggestions were judged more novel, while the Excite ratings were judged more relevant and interesting. In some cases, one system was markedly superior to the other. 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.

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

## References

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

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

[3] 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.

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

[5] 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.

[6] Ellen Spertus. ParaSite: Mining structural information on the web. In *Proceedings of the Sixth International World Wide Web Conference*, April 1997.

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