The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site

Ed H. Chi, Peter Pirolli, James Pitkow

Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{echi,pirolli,pitkow}@parc.xerox.com

ABSTRACT

Designers and researchers of users' interactions with the World Wide Web need tools that permit the rapid exploration of hypotheses about complex interactions of user goals, user behaviors, and Web site designs. We present an architecture and system for the analysis and prediction of user behavior and Web site usability. The system integrates research on human information foraging theory, a reference model of information visualization and Web data-mining techniques. The system also incorporates new methods of Web site visualization (Dome Tree, Usage Based Layouts), a new predictive modeling technique for Web site use (Web User Flow by Information Scent, WUFIS), and new Web usability metrics.

Keywords

Information foraging, information scent, World Wide Web, usability, information visualization, data mining, longest repeated subsequences, Dome Tree, Usage-Based Layout.

INTRODUCTION

The World Wide Web is a complex information ecology consisting of several hundred million Web pages and over a hundred million users. Each day these users generate over a billion clicks through the myriad of accessible Web sites. Naturally, Web site designers and content providers seek to understand the information needs and activities of their users and to understand the impact of their designs. Given the magnitude of user interaction data, there exists a need for more efficient and automated methods to (a) analyze the goals and behaviors of Web site visitors, and (b) analyze and predict Web site usability. Simpler, effective, and efficient toolkits need to be developed to explore and refine predictive models, user analysis techniques, and Web site usability metrics.

Here we present an architecture and system for exploratory data analysis and predictive modeling of Web site use. The architecture and system integrates research on human information foraging theory [6], a reference model of information visualization [3], and Web data-mining techniques [9]. The system also incorporates new methods of Web site visualization, a new predictive modeling technique for Web site use, and new Web usability metrics. The system is currently being developed for researchers interested in modeling users within a site and investigating Web site usability; however, the ultimate goal is to evolve the system so that it can be effectively employed by practicing Web site designers and content providers.

WEB SITE ANALYSIS AND PREDICTION

Most Web sites record visitor interaction data in some form. Since the inception of the Web, a variety of tools have been developed to extract information from usage data. Although the degree of reliability varies widely based upon the different heuristics used, metrics like the number of unique users, number of page visits, reading times, session lengths, and user paths are commonly computed. While some tools have evolved into products[1], most Web log file analysis consists of simple descriptive statistics, providing little insight into the users and use of Web sites.

A new emerging approach is to employ software agents as surrogate users to traverse a Web site and derive various usability metrics. WebCriteria SiteProfile uses a browsing agent to navigate a Web site using a modified GOMS model and record download times and other information. The data are integrated into metrics that assess: (a) the load times associated with pages on the site, and (b) the accessibility of content (ease of finding content). The accessibility metric is based upon the hyperlink structure of the site and the amount of content. An analysis of the actual content is not performed.

Current approaches to Web site analysis are aimed at the Webmasters who are interested in exploring questions about the *current* design of a Web site and the *current* set of users. However, Webmasters must also be interested in *predicting* the usability of *alternative designs* of their Web sites. They also seek to answer these same questions for new kinds of (hypothetical) users who have slightly different interests than the current users.

Our work aims to develop predictive models capable of simulating hypothetical users and alternative Web Site design. Using these models, we also seek to develop means for the automatic calculation of usability metrics. Our

CHI '2000 The Hague, Amsterdam
Copyright ACM 2000 1-58113-216-6/00/04...$5.00

[1] For instance, Accrue Insight (http://www.accrue.com), Astra SiteManager (http://www.merc-int.com), and WebCriteria SiteProfile. (http://www.webcriteria.com).

research on new analysis models, predictive models, and usability metrics contribute to the development of tools for the practicing Web site designer interested in exploring "what-if" Web site designs.

Our system was developed to answer questions beyond those answered by basic descriptive statistics. Specifically, we sought to answer questions concerning the entire Web site, specific pages, and the users:

- *Overall site.* What is the overall current traffic flow? What are the actual and predicted surfing traffic routes (e.g., branching patterns, pass-through points)? How does the site measure on ease of access (finding information) and cost?
- *Given page.* Where do the visitors come from (i.e., what routes do they follow)? Where do they actually go? What other pages are related?
- *Users.* What are the interests of the visitors (real or simulated) to this page? Where do we think they should go given their interests? Do actual usage data match these predictions, and why? What is the cost (e.g., in terms of download time) of surfing for these visitors?

INFORMATION FORAGING AT WEB SITES

Information foraging theory [6] has been developed as way of explaining human information-seeking and sense-making behavior. Here we use the theoretical notion of *information scent* developed in this theory [5,6] as the basis for several analysis techniques, metrics, and predictive modeling. We also employ a data mining technique involving the identification of *longest repeated subsequences* (LRS, [9]) to extract the surfing patterns of users foraging for information on the Web. This fusion of methods provides a novel way of capturing user information goals, the affordances of Web sites, and user behavior.

Information Goals and Information Scent

On the Web, users typically forage for information by navigating from page to page along Web links. The content of pages associated with these links is usually presented to the user by some snippets of text or graphic. Foragers use these *proximal* cues (snippets; graphics) to assess the *distal* content (page at the other end of the link).[2] Information scent is the imperfect, subjective, perception of the value, cost, or access path of information sources obtained from proximal cues, such as Web links, or icons representing the content sources.

In the current system, we have developed a variety of approximations for analyzing and predicting information scent. These techniques are based on psychological models [6], which are closely related to standard information retrieval techniques, and Web data mining techniques based on the analysis of Content, Usage, and hyperlink Topology (CUT, [3,10]). For more details, see [1].

Reverse Scent Flow to Identify Information Need

A well-traveled path may indicate a group of users who have very similar information goals and are guided by the scent of the environment. Therefore, given a path, we would like to know the information goal expressed by that path. We have developed a new algorithm called Inferring User Need by Information Scent (IUNIS) that uses the Scent Flow model in reverse to determine users' information goals [1]. Such goals can be described by a sorted list of weighted keywords, which can be skimmed by an analyst to estimate and understand the goals of users traversing a particular path.

Mining Web Site Foraging Patterns

Pitkow and Pirolli [9] systematically investigated the utility of a Web-mining technique that extracts significant surfing paths by the identification of longest repeating subsequences (LRS). They found that the LRS technique serves to reduce the complexity of the surfing path model required to represent a set of raw surfing data, while maintaining an accurate profile of future usage patterns. In essence, the LRS technique extracts surfing paths that are likely to re-occur and ignores noise in the usage data. We use the LRS data mining technique to identify significant surfing paths in real and simulated data.

Overview of the Analysis Approach

Our assumption is that, for the purposes of many analyses, users have some information goal, and their surfing patterns through the site are guided by information scent. Given this framing assumption we have developed techniques for answering a variety of Web site usability questions. First, for a particular pattern of surfing, we seek to infer the associated information goal. Second, given an information goal, some pages as starting points, and the information scent associated with all the links emanating from all the pages, we attempt to predict the expected surfing patterns, and thereby simulate Web site usage. Finally, we develop metrics concerning the overall goodness of the information scent that leads users to goal content (cf. [11]). Using these methods, we analyze the quality of Web links in providing good proximal scent that leads users to the distal content that they seek.

ARCHITECTURE

The architecture of the system is based on the Information Visualization Reference Model [3]. Figure 1 shows the architecture of the system using the Data State Model and the associated operators. The figure summarizes the data states and the operators defined by the system components, which we describe in detail below.

In this figure, circles represent the data states, while edges represent operators. There are four major data state stages: Value, Analytical Abstraction, Visualization Abstraction, and View. There are three major operator types: Data, Visualization, and Visual Mapping Transformations. The right side of the figure depicts these stages and types.

[2] Furnas referred to such intermediate information as "residue" [4].

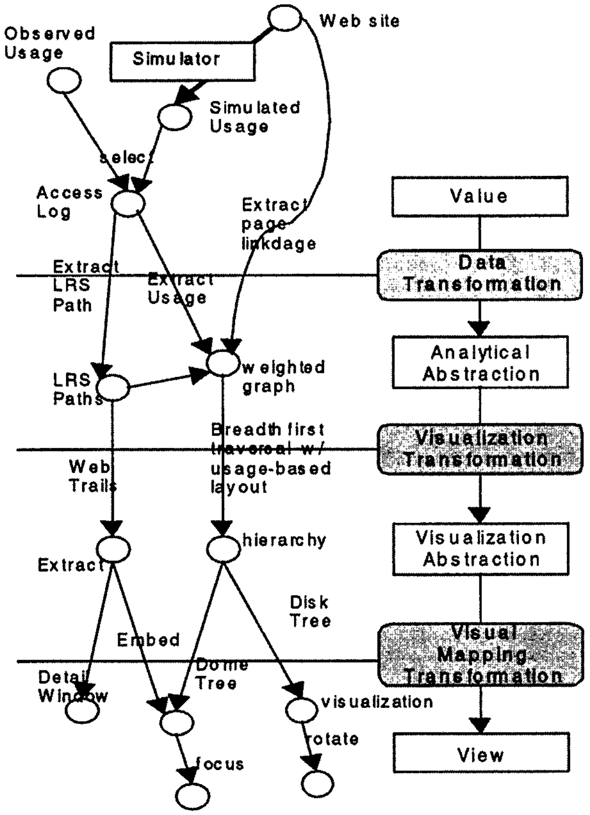


Figure 1: Data State Model for Web Scent Visualization

At the conceptual level, Figure 1 shows an important feature of the architecture: the actual observed usage data can be seamlessly replaced by simulated usage data, without disturbing other parts of the system. By pushing the observed or the simulated surfing data through the system, we obtain visualizations of actual or simulated usage. By providing this capability, users of the system can quickly test hypothetical cases against actual usage in a real-time, iterative manner, thus supporting detailed investigation into a site's usability.

SYSTEM FOR WEB SCENT VISUALIZATION

Using the reference model, we constructed a system for visualizing and analyzing a site's information scent, user trails, and usability. In the next sections, we describe the system components, followed by a series of cases illustrating the utility of the system.

Web Site and Observed Data

To develop and test the system, we used data collected at www.xerox.com on May 18th, 1998. Although slightly dated, the data set has been explored for a variety of other purposes [8,9] and was chosen to enable cross study comparisons and validation. The snapshot consists of roughly 15,000 pages and its associated Content, Usage, Topology (CUT) data. Content and topology data were extracted from the actual Web site using the techniques outlined in [7]. Usage data were extracted from the Extended Common Log Format access logs using the Timeout-Cookie method [8] to identify individual paths of contiguous surfing of Web pages by individual users.

Simulated Data

For the simulated data we developed a new technique called *Web User Flow by Information Scent* (WUFIS) [1]. Conceptually, the simulation models an arbitrary number of agents traversing the links and content of a Web site. The agents have information goals that are represented by strings of content words such as "Xerox scanning products." For each simulated agent, at each page visit, the model assesses the information scent associated with linked pages. The information scent is computed by comparing the agents' information goals against the pages' contents. This computation is a variation of the computational cognitive model for information scent developed in [6]. The information scent used by the simulation may be the distal scent of the actual linked content, or the proximal scent of the linked pages as represented by a text snippet or icon. For the cases examined in this paper, we used simulations based on the distal information scent, but, as we shall illustrate, this turns out to be fruitful way of identifying problems with the way pages are presenting proximal information scent.

Usability Metrics

We are developing metrics to assess the quality of scent at a Web site in leading users to information they are seeking, and the cost of finding such information. One of these metrics involves (a) the specification of a user information goal (e.g., "Xerox products"), (b) the specification of one or more hypothetical starting pages (e.g., the Xerox home page), and (c) one or more target pages (e.g., a Xerox product index). Using the WUFIS simulator, agents traverse the Web site making navigation decisions based on the information scent associated with links on each page. The navigation decisions are stochastic, such that more agents traverse higher-scent links, but some agents traverse lower-scent links [1]. The simulation assumes that the agents either stop at the target page when found, or failing to find the page they surf up to some arbitrary amount of effort. We then assess the proportion of simulated agents that find the target page.

Network Representations of CUT

CUT graphs and various derivatives are readily extractable from most Web sites and the corresponding usage logs. In this representation, nodes in the graph correspond to Web pages and weighted directed edges correspond to the strength of association between any pair of nodes. For the analyses in this paper, we extracted the following graphs:

- *content similarity* graph [7], represents the similarity between Web pages as determined by the textual content of the pages. The edge values provide an approximate measure of the *topical relevance* of one page to another.
- *usage* graph [7], represents the proportion of surfers that traverse the hyperlinks between pages. The edge

values reflect how users "voted with their clicks" in finding relevant information.

- *co-citation* graph [10], reflects the frequency that two nodes were linked to by the same page. The edge values provide an indication of the *authoritative relevance* of pages to one another.

Spreading Activation Assessments of Scent

We use a spreading activation algorithm [7] on the various graphs to compute relevance or scent over a Web site. Conceptually, spreading activation pumps a metric called activation through one or more of the graphs. Activation flows from a set of source nodes through the edges in the graph. The amount of activation flow among nodes is modulated by the edge strengths. In this model, source nodes correspond to Web pages for which we want to identify related pages. After a few iterations, (subject to the selection of the appropriate spreading activation parameters), the activation levels settle into a stable state. The final activation vector over nodes defines the degree of relevance for a set of pages to the source pages.

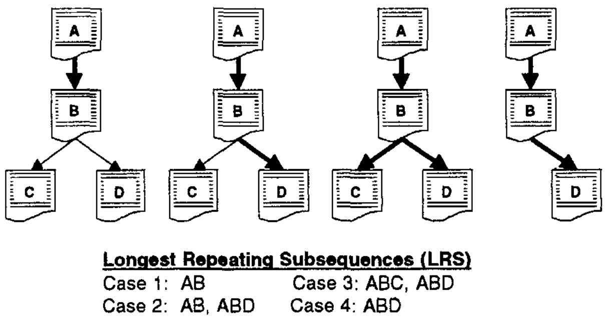
Surfer Patterns Identified by LRS

A longest repeating subsequence (LRS) is a sequence of items where (1) subsequence means a set of consecutive items, (2) repeated means the item occurs more than some threshold *T*, where *T* typically equals one, and (3) longest means that although a subsequence may be part of another repeated subsequence, there is at least once occurrence of this subsequence where this is the longest repeating.

To help illustrate, suppose we have the case where a site contains the pages A, B, C, D, where A contains a hyperlink to B and B contains hyperlinks to both C and D. As shown in Figure 2, if users repeatedly navigate from A to B, but only one user clicks through to C and only one user clicks through to D (as in Case 1), the LRS is AB. If however more than one user clicks through from B to D (as in Case 2), then both AB and ABD are LRS. In this event, AB is a LRS since on at least one other occasion, AB was not followed by ABD. In Case 3, both ABC and ABD are LRS since both occur more than once and are the longest subsequences. Note that AB is not a LRS since it is never the longest repeating subsequence as in Case 4 for the LRS ABD.

Dome Tree Visualization

Chi et al. [2], developed a visualization called the Disk Tree to map large Web sites. See the right side of Figure 3 for an example. At the center of the Disk Tree is the root node, and successive levels of the tree are mapped to new rings expanding from the center. The amount of space given to each sub-tree is proportional to the number of leaf nodes it contains.



<u>Longest Repeating Subsequences (LRS)</u>
Case 1: AB Case 3: ABC, ABD
Case 2: AB, ABD Case 4: ABD

Figure 2. Examples illustrating the formation of longest repeating subsequences (LRS). Thick-lined arrows indicate more than one traversal whereas thin-lined arrows indicate only one traversal. For each case, the resulting LRS are listed.

One of the limitations of this approach is that overlaying user paths on top of the Disk Tree occludes the underlying structure of the Web site, removing important visual data from the analyst's view. With our current focus on the flow of users through web sites, we designed a new technique called Dome Tree. In a Dome Tree, only 3/4 of the disk is used and at each successive level, the disk is extruded along the Z dimension. The rationale behind using extrusion is expanding the structure to 3D so that we can embed user paths in 3D rather than on the surface of the Disk Tree. By using only 3/4 of the disk, we can peer into the Dome through the opening like a door, without being occluded by the object itself. While this provided a useful layout, we sought to further minimize the impact of path crossings inherent in visualizing Web trails.

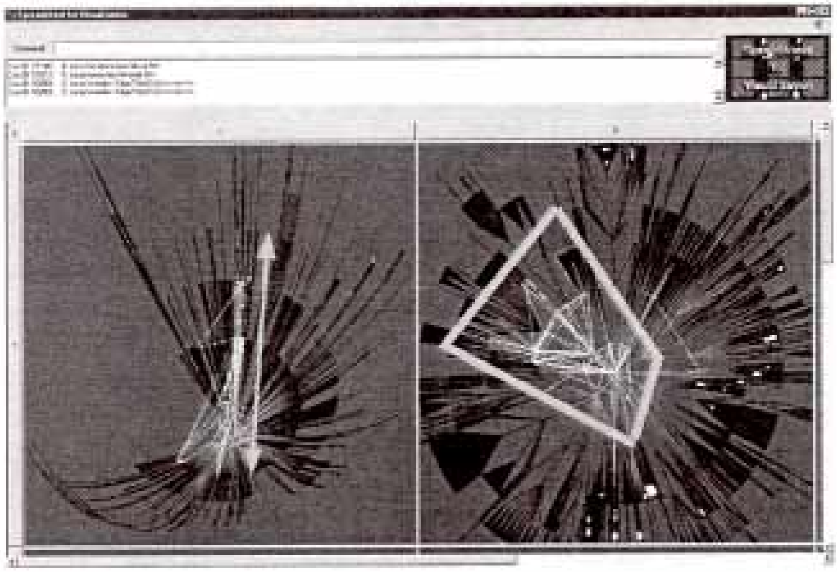


Figure 3: Dome Tree with Usage-Based Layout (left) shows that links (shown in yellow) are laid along significant paths (shown by orange arrow), eliminating crossings. In comparison, the traditional Disk Tree approach (right) has many crossing yellow links (shown in enclosed orange box). White arrows point to the current document beina examined (investor.html).

Usage-Based Layout

To provide a visualization of Web paths with less path crossings, we developed new layout methods called Usage-Based Layout (UBL). UBL algorithms determine hierarchical relationships by various popularity metrics derived from user's paths and usage data. These methods represent a departure from traditional graph layout methods that rely exclusively upon the traversal of structural relationships.

Applied to the Web, UBL can also identify user paths between two pages even though no explicit hyperlink exists between the two pages. We call this *link induction.* Link induction finds usage paths that arise from the use of history buttons, search result pages, other dynamic pages, and so forth, which cannot be obtained by crawling the site.

To determine the hierarchical relationships between documents, we conduct a priority-based traversal based on usage data. Starting from the root node, its children are determined by looking at the existing hyperlink structure as well as the inducted links. Instead of using a simple queue as in a breadth-first traversal algorithm, we use a priority queue, where the top-most used page is chosen as the next node to expand. The expanded children list is then sorted in increasing usage order, and then inserted at the end of the queue along with their usage data. Then we proceed to the next highest-used child of the root node, which is at the top of the queue.

Figure 3 displays LRS user trails using the Dome Tree with UBL as compared with the Disk Tree. The green structure is the map of the Web site, and the yellow and blue lines represent user trails. This example demonstrates that we are able reduce trail crossings by using UBL.

By using a mouse-brushing technique, we highlight each node and show its URL and frequency of usage as the mouse moves over the documents on the Dome Tree (left of Figure 3). An orange ball highlights the current document of interest. The user is then allowed to pick a particular document to bring up additional details on it.

Web Trails

One of the details shown is the extracted Web Trails that are made by the users. All paths that lead into this document are called *History Trails*, which are shown in blue. All paths that spread out from this document are *Future Trails*, which are shown in yellow.

A dialog box also pops up, containing trail information related to this document (See Figure 4 for an example). The dialog shows the history and future portion of each path, along with its length and frequency. A scrollbar on the right enables the user to graze over this list. The bottom of the dialog box shows the documents that are on these paths, with their frequency of access, size, and URL.

Clicking on a path or a portion of a path narrows the list of documents to just the documents on that particular path. In this way, we enable analysts to drill down to specific paths of interest. Selecting a path also highlights it in the Dome Tree visualization in red.

Clicking on the Reverse Scent button in the dialog box dynamically computes a set of keywords using IUNIS that describes the information needs expressed by that path. The list is shown to the user in sorted order, with the most diagnostic words at the top.

We also compute and show an estimated download time of a user traversing this path using a modem. The estimation is derived from the total bytes of the files on the path. Analysts can therefore quickly judge the cost of traversing this path, and make appropriate judgements on the path's usability.

Scent Visualization

The user can choose to show several kinds of scent related to the selected document, including spreading activation based on Content Similarity, Co-citation, and Usage graphs, and WUFIS-computed scent flow. The system dynamically computes these scent assessments for each document and shows the result using red bar lines on the Dome Tree. The taller the red bar, the higher the scent.

By visually comparing the documents that lie on user trails and the computed scent, we can see whether users are finding the information that they need. This gives us a direct visual evidence of goodness of the design of the Web site. If the paths and the scents match, then users are navigating the Web site with success. If the paths and the scents mismatch, then it is possible that users are not finding the information because the Web site design gives inappropriate scents.

In practice, we have found Spreading Activation based on Content Similarity and Scent Flow computed by WUFIS to be very useful. Therefore, we have included this information as a column in the bottom portion of the dialog box. A mark of "C" means Content Spreading Activation predicted its relevance, and a mark of "S" means Scent Flow predicted its relevance.

Overview

This section described our system for the analysis and visualization of information scent, user surfing, and Web usability. The interactions between the different pieces of the components enable analysts to mine both the actual and predicted usage data of a large Web site. Looking at the architecture depicted in Figure 1, one important data flow through the components is
Log+Web site → LRS+Graph → Hierarchy → Dome Tree.
This path uses the Usage-based Layout to compute a Dome Tree, which visualizes the whole site, with room to accommodate the Web Trails. Another data flow is
Log → LRS Paths→ Web Trails → Embed on Dome Tree, which computes the appropriate trails that are to be embedded on the Dome Tree.
In the next section, we will show the tool in action, and present a number of case studies.

CASE STUDIES

Earlier in this paper, we presented questions that might be posed about surfers and Web sites. In this section, we illustrate the system by various analysis scenarios of the Xerox Web site[3]. Specifically, we will attempt to answer the following questions:

1. What pages act as *multi-way branching* points for user traversals? Do users branch on these pages? What pages behave as *pass-through* points?
2. For a page, what are the *well-traveled paths*? Do users find the desired information on these paths?
3. For well-traveled paths, what is the *users' information goal*? How can this information goal be extracted?
4. What are the *predicted useful information destinations*, given a specific information need? Does actual usage conform to these predictions? Why, or why not?

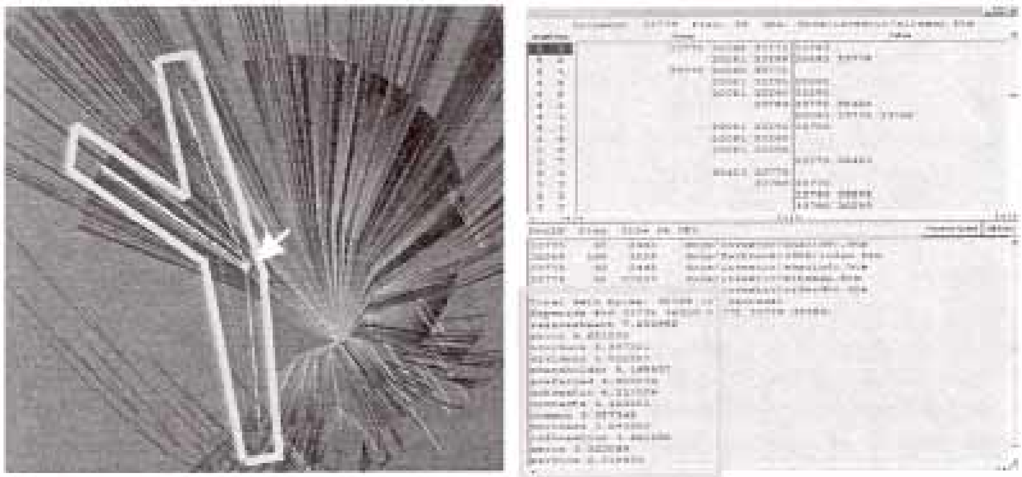


Figure 4: Multi-way Branching Point (investor/sitemap.htm) shown enclosed by orange lines, and Web Path detail dialog box (orange box shows the inferred user information need keywords, which are reinvestment, stock, brochure, dividend, and shareholder).

Page Types

Some pages act as indices, serving as *way points* in navigation patterns. Other pages act as *conduits* in a set of serially organized pages. Given these and other page types, the question arises, "How are users actually surfing these pages?" One may posit the design principle that effective way points should be kept around for good navigational scent. Once identified, ineffective way points can be redesigned, integrated with other content, or removed.

Figure 4 reveals a multi-way branching point where a few history paths lead into the branching point and result in a few well-traveled future paths. Upon drill-down, we discover that the branching page leads to several important destination pages, including the shareholder information page, the 1998 Xerox Fact Book, and a financial document-ordering page. While the page is relatively under-utilized (~60 accesses/day), our analysis shows it to be a very effective local sitemap. Within a few clicks, users are able to access the desired content.

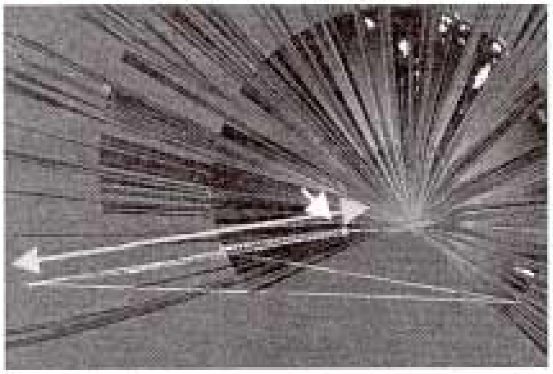


Figure 5: Pass-through Point in a series of pages (marked by orange arrows and current page pointed by white arrow is annualreport/1997/market.htm)

Figure 5 shows an example of a pass-through point where UBL has laid out the pages in path-priority order. In traversing this path, some users leave the serial organization of the pages to find a related page (yellow path going to the red Content Spreading Activation page, bottom right). Users then backtrack to continue surfing the serial links. From this inspection, we conclude that while it is a fairly well designed pass-through point, the page could potentially be improved to incorporate the related content directly. The tradeoff may be between coherence of the pages and navigational effort.

Well-Traveled Paths

Currently, most Web site visualizations focus on the identification of high usage areas. Our system identifies well-traveled paths by using a combination of two methods. First, the LRS computation reduces the number and complexity of user paths into manageable chunks. Second, embedding the paths onto the Dome Tree facilitates the visual extraction of well-traveled paths. We do not consider these methods perfect, rather they permit investigations that are otherwise difficult to attempt.

The left-hand image of Figure 6 illustrates the well-traveled paths related to a specific Web page (the TextBridge Pro 98 product page). As evidenced by the myriad of yellow future paths, related information is laid out across many different areas of the Web site, suggesting a possible redesign to bring more cohesion into the site. One

[3] Since the system is built for displays exceeding the resolution of paper, we have placed a copy of the figures for inspection online at: http://www.parc.xerox.com/uir/pubs/chi2000-scent.

interesting well-worn path is the serial pattern on the left (long arching yellow and blue path) that corresponds to the product tutorial pages. The right-hand image of Figure 6 shows the well-traveled paths extending from the Pagis product page. The zigzagging paths near the page indicate surfing between popular sibling pages. Many users travel the software demo tour and this is made explicit by the large blue path radiating upwards.

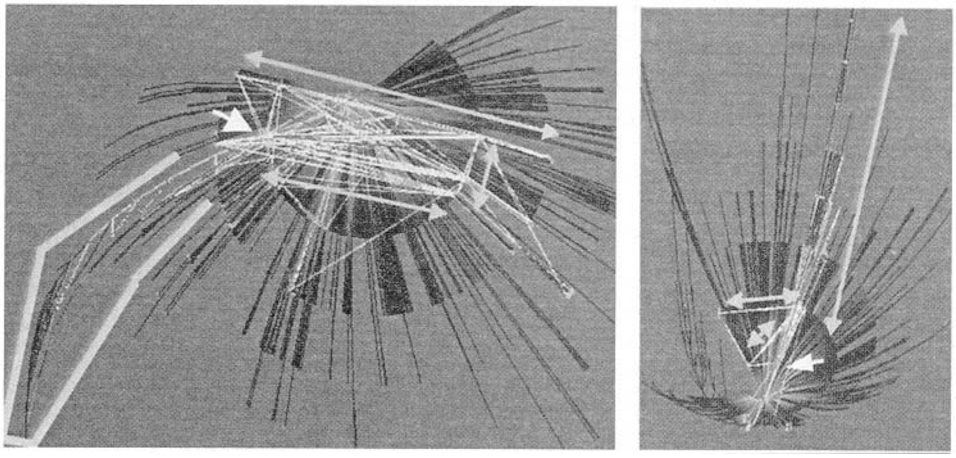


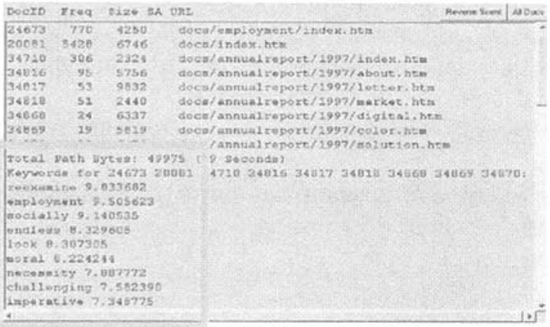
Figure 6: Well-traveled paths related to scansoft/tbpro98win/index.htm (left) and scansoft/pagis/index.htm (right), where major traffic routes are marked by orange lines.

In both of these examples, the red bar marks throughout the Dome Tree indicate the related pages to TextBridge and Pagis as computed by the Scent model. The correspondence of predicted related content to actual user paths suggests that the related content is not only reachable, but also well traveled by users. Visually, the yellow user paths that connect the red bars extending from the related pages reveal this correlation.

Identifying Information Need

Since well-traveled paths indicate items that compete well against other items for users' attention, it is important to find out, given a well-traveled path, what information need has the user expressed in that path. The bottom of Figure 4 shows the information need of a well-traveled path as computed by the reverse scent algorithm. The example is taken from a path related to investor/sitemap.htm. The top keywords computed by the reverse scent algorithm are reinvestment, stock, brochure, dividend, and shareholder. These keywords represent the goal of the users that traverse the path from the Shareinfo to the Orderdoc Web pages.

Figure 7 (corresponding to Figure 5) shows a more specific information need for the highly traversed path that starts at the employment recruitment page and winds through the 1997 Annual Report. In this case, some of the top keywords are reexamine, employment, socially, and morals, suggesting that potential Xerox employee are investigating the attitudes and culture of Xerox as expressed in the Annual Report. Another possible interpretation is that researchers are examining the correspondence between Xerox's employment policy and its social/moral position. In Figure 8, a large number of paths relate to how to upgrade previous versions of TextBridge96. A representative path shows top keywords as TextBridge, upgrade, OCR, Pro, bundled, software, windows, and resellers.



```
DocID  Freq  Size SA URL                                    Reverse Scent | All Docs
24673   770   4250    docs/employment/index.htm
20081  5428   6746    docs/index.htm
34710   306   2324    docs/annualreport/1997/index.htm
34816    95   5756    docs/annualreport/1997/about.htm
34817    53   9832    docs/annualreport/1997/letter.htm
34818    51   2440    docs/annualreport/1997/market.htm
34868    24   6337    docs/annualreport/1997/digital.htm
34869    19   5819    docs/annualreport/1997/color.htm
                          /annualreport/1997/solution.htm
Total Path Bytes: 49975 ( 9 Seconds)
Keywords for 24673 20081  4710 34816 34817 34818 34868 34869 34870:
reexamine 9.833682
employment 9.505623
socially 9.140535
endless 8.329608
look 8.307305
moral 8.224244
necessity 7.887772
challenging 7.582398
imperative 7.348775
```

Figure 7: from annualreport/1997/market.htm (Figure 5)

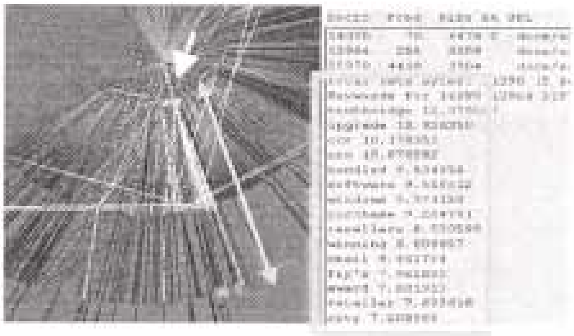


Figure 8: from xis/tbpro96win/index.htm

These examples suggest that we are able to automatically identify the information goals of users by first discovering the well-traveled paths and then computing the informative keywords using the Scent Flow model. These examples help demonstrate that the Scent Model is not only good at predicting future user surfing behavior given a starting page, but also good at determining the information needs of a set of users given their paths through a site.

Predicted Destinations Based on Scent

One analysis centers on the differences between the WUFIS Scent Flow Model and actual user behaviors. We seek to answer the question, "Where in the Web site does the Scent Flow model differ from observed data, and why?"

For example, 100 hypothetical users interested in information related to Pagis product were simulated to flow through the Web site from two different entry points. Figure 9 shows the result of these two simulations, where

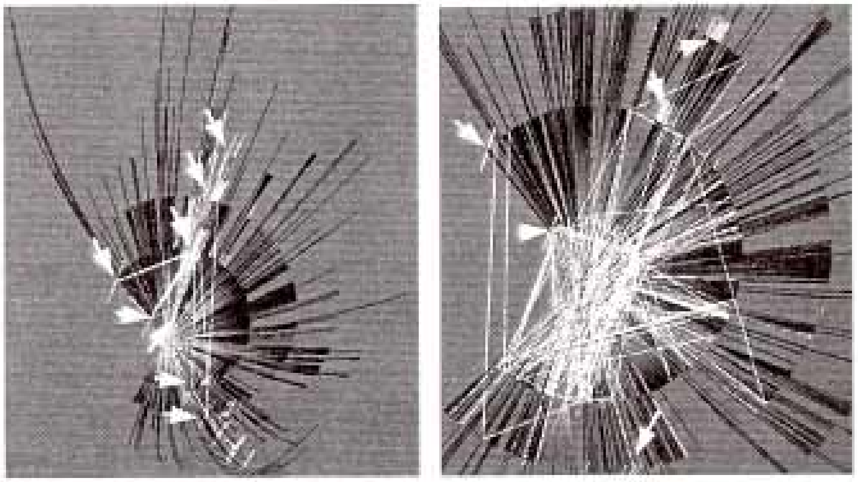


Figure 9: Given an information need related to "Pagis", Scent Flow simulation results in good match in scansoft/pagis/index.html (left, good match points pointed by orange arrows), but poor match from products.html (right, bad match points pointed to by purple arrows).

actual user paths are encoded with yellow lines and the frequency of visit by the simulated users is encoded by the height of red bars. In the left-hand image, we placed users at scansoft/ pagis/index.html, and watched the users surf to various points in the Web site, including pages relating to a tour of the software, release notes, and software registration pages. The correspondence of the yellow trails to the red pages reveals a match between the flow of real and simulated users. The right-hand image of Figure 9 displays the result of simulating users from products.html. It is immediately clear from the picture that many pages containing information relating to "Pagis" are found by the simulation, but real users are not finding pages. Upon careful examination we discovered that while the “Pagis” scent is contained near products.html, the scent is buried in layers of graphics and texts. The example shows that products.html does not adequately provide access to information relating to “Pagis”.

There remain many limitations to the current system that remain for future work. Although we have ameliorated some of the visual clutter problems associated with visualizing Web sites and user paths, there is clearly much room for improvement. Techniques such as animation might aid in showing and comparing Web Trails. Another way to improve the current visualization of Web Trails is to fade colors out as we move into history or future portion of the path. To do this, we would have to first compute the aggregate path flow down each section over all paths.

CONCLUSION

Within the last few years, we have seen an explosive growth in Web usability as a field. Given its infancy, it is not surprising that there are so few tools to assist Web analysts. We presented a Scent Flow model for predicting and analyzing Web site usability. The analysis and visualization system presented in this paper is aimed at improving the design of Web sites, and at improving our understanding how users forage for information in the vast ecology of the Web.

Acknowledgement

This research was supported in part by an Office of Naval Research grant No. N00014-96-C-0097 to Peter Pirolli and Stuart Card.

REFERENCES

1. Chi, E. H., Pirolli, P., Pitkow, J. (1999) Using Information Scent to Model Us r Information Needs and Actions on the Web. (submitted).
2. Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., and Card, S. (1998). Visualizing the Evolution of Web Ecologies. *Proceedings of the Human Factors in Computing Systems, CHI '98.* (pp. 400-407). Los Angles, CA.
3. Chi, E.H. and Riedl, J.T. (1998). An operator interaction framework for visualization systems. *Proceedings of the IEEE Information Visualization Symposium.* (pp. 63-70).
4. Furnas, G.W. (1997). Effective view navigation. *Proceedings of the Human Factors in Computing Systems, CHI '97* (pp. 367-374), Atlanta, GA.
5. Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '97* (pp. 3-10), Atlanta, GA.
6. Pirolli, P. and Card, S.K. (in press). Information foraging. *Psychological Review.*
7. Pirolli, P., Pitkow, J., and Rao, R. (1996). Silk from a sow's ear: Extracting usable structures from the web. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96* Vancouver, Canada.
8. Pirolli, P. and Pitkow, J.E. (1999). Distributions of surfers' paths through the World Wide Web: Empirical characterization. *World Wide Web, 1*, 1-17.
9. Pitkow, J. and Piroll, P. (1999, in press). Mining longest repeated subsequences to predict World Wide Web surfing. *Proceedings of the USENIX Conference on Internet.*
10. Pitkow, J. and Pirolli, P. (1997). Life, death, and lawfulness on the electronic frontier. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '97* (pp. 383-390).
11. Spool, J.M., Scanlon, T., Snyder, C., and Schroeder, W. (1998). Measuring Website usability. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '98* (pp. 390), Los Angeles, CA.

The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site (Page 161, plate 1)

Ed H. Chi, Peter Pirolli, James Pitkow

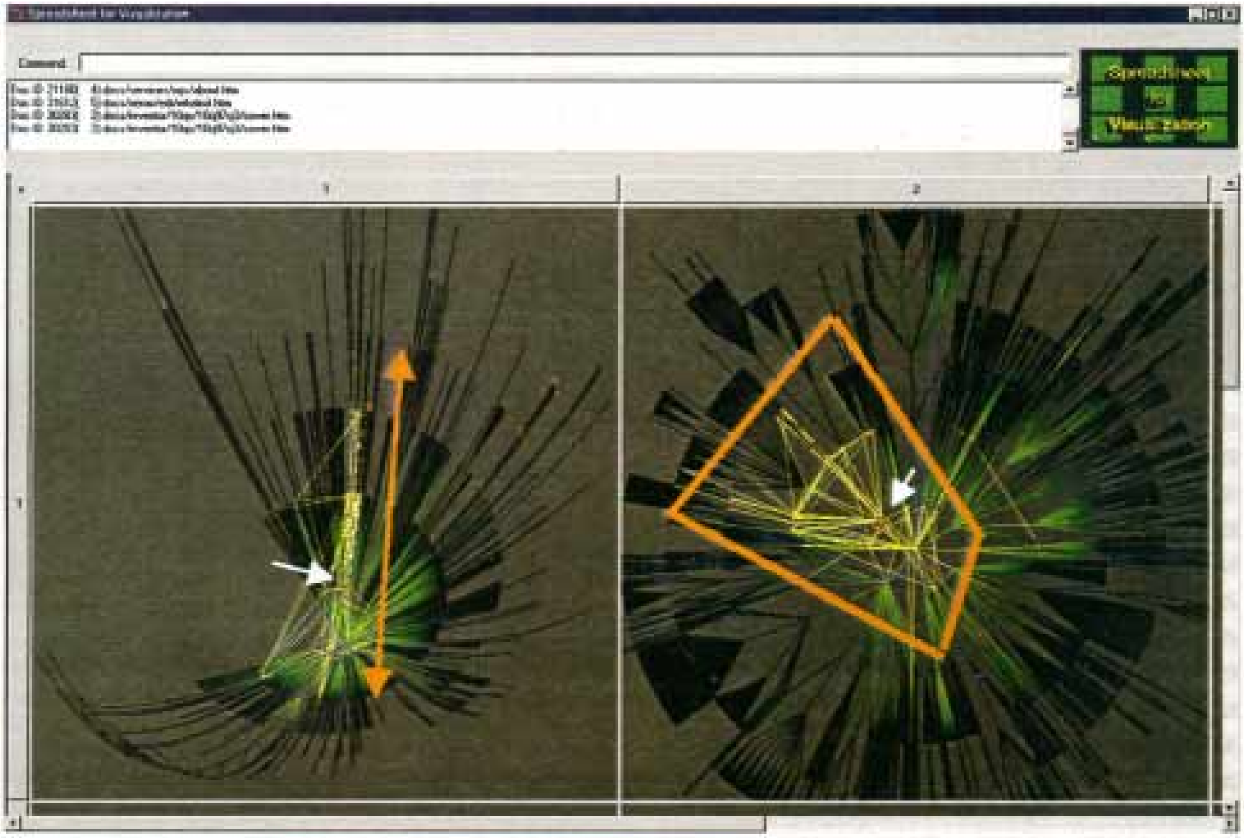


Figure 3: Dome Tree with Usage-Based Layout (left) shows that links (shown in yellow) are laid along significant paths (shown by orange arrow), eliminating crossings. In comparison, the traditional Disk Tree approach (right) has many crossing yellow links (shown in enclosed orange box). White arrows point to the current document being examined (investor.html).

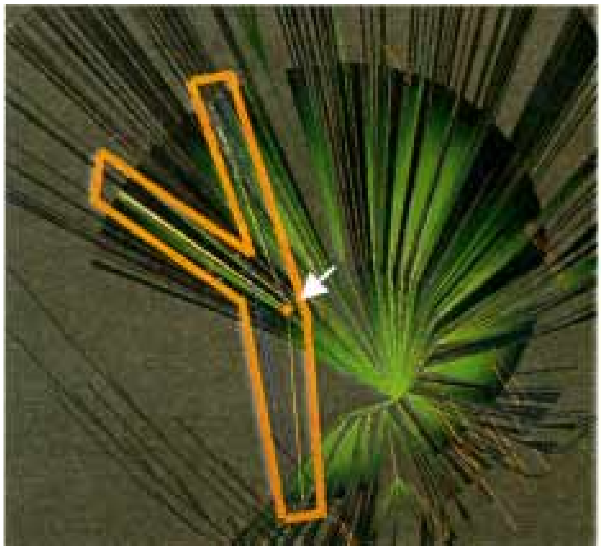


Figure 4: Multi-way Branching Point (investor/sitemap.htm) shown enclosed by orange lines.

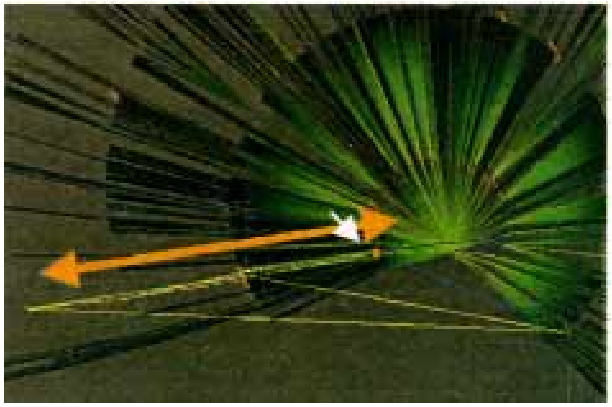


Figure 5: Pass-through Point in a series of pages (marked by orange arrows and current page pointed by white arrow is annualreport/1997/market.htm)

The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site (Page 161, plate 2)

Ed H. Chi, Peter Pirolli, James Pitko

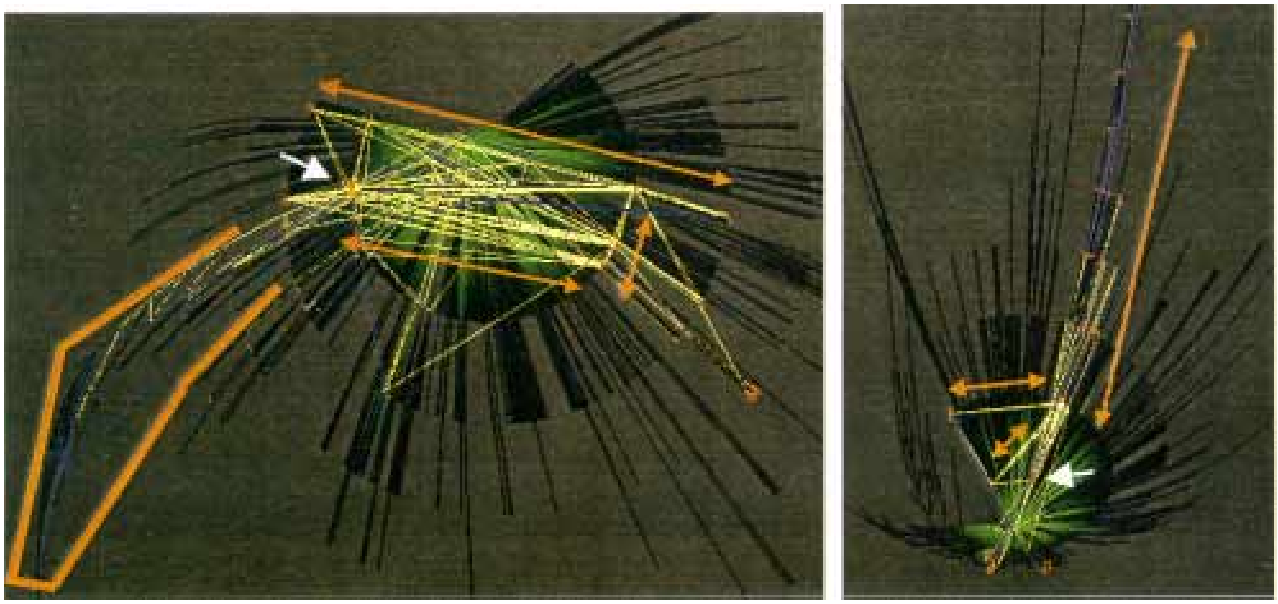


Figure 6: Well-traveled paths related to scansoft/tbpro98win/index.htm (left) and scansoft/pagis/index.htm (right), where major traffic routes are marked by orange lines and arrows.

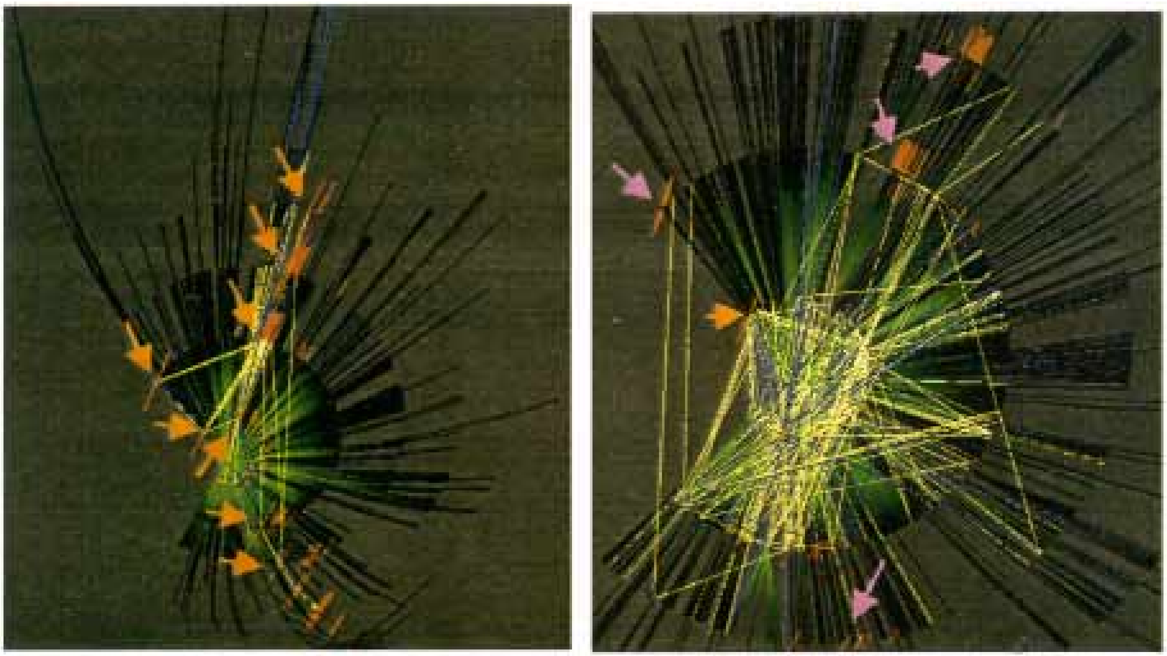


Figure 9: Given an information need related to "Pagis", Scent Flow simulation results in good match in scansoft/pagis/index.html (left, good match points pointed by orange arrows), but poor match from products.html (right, bad match points pointed to by purple arrows).

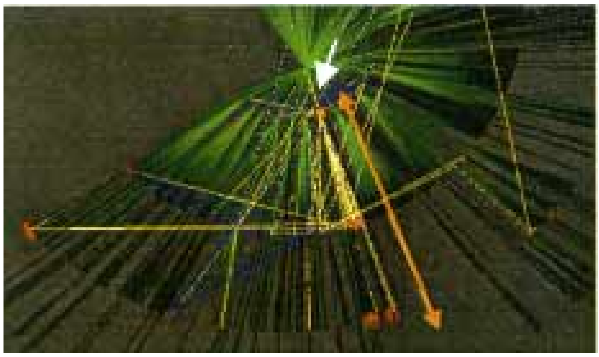


Figure 8: from xis/tbpro96win/index.htm