

Social Search Comes of Age

Deborah Richman
Senior Vice President, Collarity, Inc.
Palo Alto, CA

Abstract

For more than a decade, Information Professionals have been wrestling with how to apply effective online search mechanisms across structured and unstructured Web information. At first, Web sites were carefully identified and sorted into directories, largely by human editors. Then people were replaced by search algorithms, first driven by keywords and then based on hyperlinks among Web pages. Neither was sufficient.

We are now turning to people-powered methodologies, which may deliver the “holy grail” of personally relevant search results. This paper will define social search, how it has developed, and how it helps searchers find previously buried information on the Web.

We’ll examine different methods of social search and social sharing. While many of these social networks have value, they also often require explicit participation on the part of users – read “extra work.” We’ll examine the pros and cons of using explicit social search mechanics.

Next, we’ll examine the emergence of implicit social search participation methodologies. By taking a page from marketing segmentation and e-commerce, there are ways to deliver better search results based on collaborative community recommendations and previous behaviors. We will discuss the unique advantages and challenges of leveraging implicit user feedback.

Finally, we discuss how implicit communities can develop and contribute to relevant search results for individuals. Communities that arise from actual and live searching behaviors are reliable and valid. Individual searchers may share similar interests with expert searchers or “like minded” searchers. As a result, a richer and more relevant slice of the Web can be revealed.

Social search has come of age and there are many new available tools designed to integrate users into the relevance equation. It’s time for your patrons to understand their unique needs, choose the tools that are best suited for them, and start taking advantage of community search help.

I. The Social Search Evolution

In a world of information overload, it makes sense to leverage the collective intelligence of everyone online. Social search uses collaborative filtering to tap into this wisdom, and to produce more relevant search results.

A. Helping each other find resources

Social search may be defined as any system that enables people to help others filter and find information on the Web. It's simply an extension of the natural patterns of how humans interact with each other and with information in the real world.

Natural patterns are both explicit and implicit. Explicit interactions depend on people who actively share their opinions, including labeling or classifying Web pages for the benefit of others. Implicit interactions happen when people consume what others have contributed before. An example would be when they see recommendations such as "users who read this article, also read this article."

True social search emerged quite recently, based on the increasing bi-directional nature of the Web. In the earliest days, the Web operated as a one-way system where visitors exclusively pulled information downstream. Now, it has evolved and expanded to two-way conversations, which also depend on visitor feedback upstream. Thus, Web visitors are both consumers and producers of information.

The vision of social search is largely a democratic one. We assume that all Web visitors should be capable of helping to classify, organize, and rank the world's Web pages. They might also have expertise or interest in specific subjects. Thus the challenge is how to organize user-generated feedback so the "right people" are providing the "right answers" to the "right questions."

B. Libraries, first and foremost

In some ways, social search began long ago. Librarians helped us find the information we needed when we went looking for it. Before the Web, information professionals shared resources among peers as well as patrons. Classification systems, such as the [Dewey Decimal Classification](#) have always made this possible and have served us well for decades. Librarians determined where published information belonged on library shelves, and patrons relied on their expertise.

If you ever worked at or attended a large university years ago, you may also be familiar with a nearly extinct breed; expert cataloguers who knew their particular subject matter best and were proud of the time and effort spent organizing information for easy access. The assumption was that patrons would benefit from this early example of keyword and description tagging from experts.

C. Private online databases

By the mid-1980s, most libraries began offering electronic resources that were not available in their own holdings, including news, journals and other reference tools. Proprietary online services, like [Dialog](#), provided searchable expert abstracts, tagging and unique taxonomies depending on the reference databases shared. For example, [D&B](#) databases contained their proprietary company information that could be searched by unique fields including company revenue, employee count, addresses, year founded, and expanded industry codes.

Most early online content databases were difficult to use and still required assistance from librarians. Not everyone wanted to learn about service files and how to search them uniquely. [Lexis/Nexis](#) was (and is) a well-known vendor who made it easier for patrons to conduct their own searches, based on searching content, dates and sources. This seemed more intuitive than databases previously available to librarians.

D. Web-based libraries

Nearly a decade ago, 24/7 library access became widespread for academic, public and special libraries. The ability to search when library doors were closed became reality, as long as searchers had passwords to access the information. This new era was exciting, as patrons could search both print and electronic holdings anytime. There was no live assistance for the first time, which represented the proverbial “passing of the torch” from information professionals to patrons.

These were one-way systems where patrons consumed information from libraries. There were no upstream mechanics for providing feedback about data quality or augmenting information gaps. The social environments were defined by librarians and intermediaries who, through their holdings, assessed what was ultimately available and important to patrons. The social interplay was still fairly limited.

E. Open Web directories

During the mid-late 1990s, the explosive growth in Web sites made it difficult to find relevant search results. For the first time, anyone could set up a site and begin publishing about specific subjects of interest. It became very challenging to sort through traditional Web information as well as the non-traditional, non-expert sources that proliferated.

The earliest organizers of Web content, such as [Yahoo!](#) and [LookSmart](#), sought to provide expert guidance for whatever could be discovered online. At a basic level, they created a feedback loop for Web site publishers. As a Webmaster, you could submit your site’s information to be included in their directories. If you were excluded, then it was quite difficult to be discovered by end users. These companies and others offered a mix of searching options through directories, catalogs and reviews.

The early open online directories created new and elegant taxonomies, which grew into tremendous new resources that could put John Dewey to shame. They organized each Web site into its logical place in the hierarchy much like their old librarian counterparts. Sizeable armies were hired to do the work. This included information professionals and liberal arts graduates who spent their days working with the latest sites and manually categorizing them.

Self-selected experts volunteered to categorize sites as well. Not long ago, when you set up a new site, you submitted it for inclusion in the first open-sourced online directories and searches. These experts continued to hold sway with directories, and later search engines, for many years. Some of these open-sources, like [dmoz](#), still exist today.

Finally, there were other, more pinpointed efforts to find what was important on the Web. Expert guides were developed at [About.com](#) and written by a vast network of subject matter experts. These individual editors contributed content and also focused on identifying appropriate sites and content within their subjects.

F. The rise of search engines

A more automated form of Web content discovery came with the emergence of early search engines like [Excite](#), [WebCrawler](#), and [Lycos](#). These engines collected Web pages via “spiders” that followed all available hyperlinks to new information. When a new page was found it was added to the index, essentially a big catalogue. Each search engine developed its own proprietary algorithms to let searchers query its index.

Google then revolutionized search quality by introducing its [PageRank](#) algorithm. PageRank essentially set up a way for Web pages to “vote” for other Web pages by linking to them. For Google, sites are valued based on the sheer quantity of links to your pages. Relevancy means your page is more likely to rise to the top of the search results list or at least to the first page.

In addition to Google, many other search engines adopted similar techniques for ranking content. When you think about it, link popularity is actually one of the earliest forms of people-powered search. People or Webmasters built hyperlinks to other people’s pages. Both sides agree to cooperate as well. Links became a form of currency, with value tied to the search results.

G. Social layer needed

What is driving the need to extend search into the social realm, where patrons have a say in what information is relevant and what is not?

First, Google’s approach has been compromised by people who have figured out how to game the system. Tricking search engines and manipulating perceived relevancy has become a worldwide business! Unfortunately, search relevance based on link popularity has never been more accurate than the day it was introduced.

The continued growth in Web sites means that many valuable pages will never be found. These buried pages need to be uncovered more easily. We've seen progress in networks and machine-learning. Now the process calls for increased doses of human intelligence to help solve the information organization challenges.

Social search would add a "layer" of intelligence on top of traditional search results, thus enabling users to contribute to the relevance of search results directly. The market has simply been waiting for a few technology ingredients to catch up.

II. The Rise of Explicit Online Social Sharing

Explicit sharing simply means extending many of our social activities – conversations, labeling information, rating information, or organizing information – onto the online network. When it relates to search mechanics, someone is overtly reacting to a question or problem and is connecting to this knowledge.

Networks facilitate millions of conversations and interactions that probably wouldn't have happened if networks didn't exist. People all over the world, not just the people in the room or neighborhood, can benefit from the learning in these discussions. With networks in place, anyone online can band together to advance knowledge too. Here are examples which show how explicit sharing technologies have evolved.

A. Social bookmarking

Most people already know how to bookmark Web sites and pages through their browser. Social bookmarking is a similar concept, which allows you to save and share anything online.

Some of the more well-known sites are [Del.icio.us](#), [Furl.net](#), [Flickr](#) and [Photobucket](#)— all launched in 2003-2004. By offering a way to bookmark and tag online materials, these services continue to help users save items privately and also share findings publicly with others. Any Web pages can be shared through the first two services, while photos are mainly shared through the latter two services.

These sharing services operate as free subscriptions. After users sign up, they may begin saving and sharing Web pages immediately. Users can even annotate pages exactly like catalogue librarians did years ago with print materials. In practice, most users do far less, saving Web pages to a few tags or labels that relate to the pages.

Social bookmarking sites are valuable because humans are again in charge of defining what appeals to them independently. Their tags also get shared and help contribute to [folksonomies](#), which are user-generated taxonomies that aid future searching. Thus a valuable page that ranks poorly with the search engines may become "findable" because a group of users identified and shared it.

The disadvantage of social bookmarking systems is that they require a certain amount of training, understanding, and work to use them. Only a small minority of users takes the time to tag information and therefore some searches may reflect the explicit interests and biases of a small proportion of users. Also, some people spam results by attaching inaccurate tags to irrelevant content. This form of spamming is still difficult to control.

B. Ranking and voting

On the Web, ranking and voting activities are commonplace. Since the first consumer sites appeared, there have been polls that anyone could participate in and then click for results. These were considered a way to draw people into sites and have them understand the opinions of others who visited. While everyone who voted understood the bias, the feedback provided entertainment value.

More practically, people often like to make purchasing decisions based on the feedback from previous buyers. Ranking and voting became a force on the Internet when many retail sites began including purchaser feedback along with product information. Perhaps the most active and comprehensive rankings began appearing on comparison sites like [Shopping](#), [Shopzilla](#) and [Nextag](#); as well as on large retailer sites (see [Internet Retailer](#) rankings). For years shoppers have been asked to rank products and suppliers providing a strong search filter.

[eBay](#) may be called a pioneer in creating community-based reputation system by encouraging all buyers and sellers to provide positive or negative rankings. These are shared, in plain view, for anyone else doing business after these transactions are completed. Feedback loops create a self-policing system.

Today, Yahoo takes advantage of ranking and voting attributes via [Yahoo! Answers](#), its large scale Q&A service. When someone posts a question, he can set a deadline for the answer. Then other participants may respond to this question rather than hired experts. Ultimately the questioner may select the best response himself or open it to community voting. When searching the Q&A database, visitors may query it based on the question status, rankings and other criteria.

[Digg](#) is another example where anyone may submit articles, videos or podcasts they see on the Web. These items become more popular if others “digg” them and less popular when “buried” by users. The search reflects these attitudes and results can be sorted by most popular diggs, date or relevancy. As an open voting system, there have been issues raised about some users who game the results. There are primary benefits to the articles and sources which have high “digg” votes – and are more visible through the current search engines like Google.

The advantages of voting and ranking systems are the fact that they are ostensibly democratic and the results can often be instructive if enough community members vote. Like any voting, however, the results are vulnerable to compromising interests. Many

times the systems have difficulty tracking the online equivalent of “ballot stuffing” from organized armies of voters or single voters able to vote many times.

C. Blog discussions

Another social phenomenon occurs on active blogs. Here the blogger posts articles or opinions, and readers quickly send comments back. Depending on the interaction, blogs can become very rich resources.

Almost anyone can set up a blog and begin interacting within a wider social network based on content sharing. The rapid adoption of blogs since 2004 may be attributed to their ease-of-use and free access, using hosting companies such as [SixApart](#), [WordPress](#) or dozens of other blog hosters.

To improve searchability, bloggers actively tag their own entries based on what they believe are the most important or relevant terms. Through blog search providers like [Technorati](#), anyone may search by tagged term, blog posting, or overall blog directory. The advantage of blogs is that many new talented voices can be heard around the world. People whose opinions and knowledge would not normally be syndicated through mass media channels can now find an audience. What’s more, these social discussions can be found and accessed through search engines.

The disadvantage of blogging is, again, abuse of the system. Blogging can be the source of defamation and misinformation. Many times it is difficult to validate the claims or facts of some bloggers. The community is, however, diligently self-policing.

D. Social networking

There are a growing set of Web sites focused on making connections and sharing common interests. The most popular ones today are [MySpace](#), [Facebook](#) and [YouTube](#). All of these sites have experience phenomenal growth in the last two years.

Safa Rashtchy, who covers search at Piper Jaffray, made the point that social networks are a mix of communications and entertainment. In a [recent interview](#), he explained that active users of these social sites consider sending music files fun. This is different than earlier generations who consider these activities separately.

At sign-up, these networks all require users to establish some form of identity. They then encourage users to communicate from there. For example, YouTube asks users to select a channel: YouTuber; director; musician; comedian; or guru. It also asks for preferences regarding whether to accept comments from friends or everyone so privacy is user-defineable.

After sign-up, users can engage in reaching out to old friends and meeting new ones who share their interests. As an individual user, there are a lot of interests to discover. It’s

possible to wander around the site and find video, music and more. There are many places to browse, subscribe and search for content and people.

As a contributor, there are numerous ways to share as well. Users may comment, upload videos, and communicate with others directly. Depending on the social network, there are differences in what users can share through their account, and how they're able to define themselves and their digital persona online. For example, a user might want to be a musician on YouTube and simply an old college friend with many interests while on Facebook.

Social networks definitely provide a valuable resource for many people to stay in touch with their friends and to meet new people they otherwise might not have met. Sometimes people can connect with experts in a given subject area when they need advice or help. Much of the value comes from entertainment too.

However there are some downsides from social networks, including many instances of abuse related to users misbehaving or using fake online identities to deceive others. There have also been other cases where personal information has been divined through online conversations, leading to crimes and problems offline.

So far, this paper has presented social search that depends solely on explicit interactions. While these interactions can improve search and content discovery experiences, there are intractable issues about participation, gaming, and skewed results that simply can't be ignored. These issues impact search relevancy. Thus, implicit approaches will be explored as a way to address them.

III. The Emergence of Implicit Online Social Sharing

When using implicit systems, the main distinction is that online visitors are never asked outright about their opinions. Instead, their attitudes, tastes and preferences are captured or implied from online behaviors and activities. The old saying that "actions speak louder than words" is true here. Some systems require actual sign-ups, while others can deliver social searches based on anonymous navigations.

A. Voting with your mouse

Implicit data collection begins when a person starts clicking around a site. This data is often used to narrow Web pages to those most likely to interest that particular person. For example, if a user searches on the word "java" and subsequently looks at pages related to coffee, and not computer programming, it could be inferred that this person should receive coffee related pages when he searches on the word java the next time. Personalization gives search engines guidance on which results to emphasize and which results to filter out.

In the simplest terms, search engines can use the items and pages users click on to paint a picture of a given user's tastes or preferences. This effectively creates a profile reflecting the user's interests. This profile can be used as an additional set of information to filter search results.

Some fine tuning is involved, since not all tracked behaviors are related to a user's interests. People may mistakenly choose a page or they may go to a page that has nothing to do with subjects that interest them. Often, a certain amount of unrelated "noise" may be introduced into the system. However, given enough data, systems can be tuned to recognize a person's true preferences over time and any aberrations can be filtered out.

B. Implicit benefits

Implicit sharing creates an excellent, high quality feedback mechanism. Users aren't asked to do extra work like tagging, rating, or bookmarking. Implicit systems take advantage of the dialogue already happening between a Web site and its audience, rather than burdening individuals with unfamiliar, disruptive ones.

All users or visitors to an enabled site participate, in contrast to the minority of people willing to tag or rate actively for the benefit of others. This creates a much richer data set. An implicit system is more efficient and capable of operating in real-time, because it doesn't rely on users themselves to initiate feedback.

Implicit social search systems are just beginning to emerge. There are significant benefits to the users of these systems. There's greater search accuracy based on actions. Since implicit feedback systems gather a more complete data set, there is no survey bias, and the search results and recommendations tend to be more accurate.

By definition, implicit systems are fully distributed and used by 100% of visitors to a publisher's site. That means it is difficult for the actions of a few to influence the results, and there should be far less abuse and spam – if at all.

C. Building communities

All this implicit data should help publishers gain new perspective about the interests of their visitors, as they are able to observe hundreds of unique communities. These communities are little bit like "intelligence farms." They reflect people with common interests, along with these attributes:

- *Segmented:* Groups of individuals with the experience and knowledge to provide the best advice on specific subject areas.
- *Organic:* Communities which form and dissolve based on "hot spots" of interest somewhere on the Web. There are no artificial taxonomies.

- *Invisible*: Members who may not be aware of other community members. In addition, they are often unaware of their own membership.
- *Dynamic*: Membership which rises and falls based on activity. Communities form from a “critical mass” of interest, but dissolve when interest falls off.
- *Meritocracy*: Individuals who are “joined to” a community based on their contribution value and behavior, such as expertise or interest in a subject area.

D. Communities Matter

For Web publishers and their visitors, communities matter. The basic notion of the Web is that every possible community can be served or discovered. Site publishers have entered all kinds of vertical markets and appealed to many targeted audiences. Even within a market, there are endless subjects and interests. Publishers are challenged to define and respond to all these interests and are increasingly turning to companies like [Collarity](#) for site search and content discovery.

Implicit communities naturally mirror all active interests, via searching and browsing behaviors. Someone may want to solve a problem, research something newsworthy, address an immediate need, pursue a hobby or passion, get some work done, go shopping, or finish some repetitive task. Within a site, there could be differences among someone visiting for the first time, occasional visitors, and frequent returnees.

Communities are influenced by their visitors, but are no longer dependent on knowing exactly who’s present. For example, grade schoolers doing homework differ from high schoolers, college students or adults. Other demographics may come into play too, as interests vary depending on family status, income, education, geography and other factors.

The communities are important because they can respond to any individual, even anonymously. Using collaborative filtering, the searcher will receive guidance from different communities as needed. Thus communities are critical for improving search relevance. Without them, everyone would still see the same, one-size-fits-all results.

In the end, visitors have much greater control and flexibility through implicit sharing approaches. They can personalize their search results, get advice from communities of experts and like-minded searchers, and see results impacted by all users of the system.

IV. Summary: Power to Consumers

The most logical way to advance search relevance is for search engines to give users the opportunity to participate more broadly in the system. Explicit forms of information sharing suffer from survey bias and low participation rates. Emerging implicit social sharing systems hold the potential to increase accuracy and control for users.

During 2007, the outlook looks very strong for social search overall. That's due to the confluence of increased social computing activity by end-users, expanded demand by publishers and new capabilities from social system providers.

End users have gotten accustomed to interacting online. In addition to reading, listening or viewing activities, they are spending time sharing and communicating too. While it's well known that college students and teenagers spend substantial online time engaged with others, recent research shows that adult participation has reached mainstream levels as well.

Forrester Research ([April 2007](#)) reported that 48% of U.S. adults participated in some form of social computing. Some 13% have uploaded content, 19% posted comments, 15% bookmark and tag, and 19% participate in social communities. Forrester found that 33% are spectators or the audience for user-generated content such as blogs, podcasts and videos.

Pew/Internet American Life Project ([May 2007](#)) looked at online and cell phone users. Nearly 19% of U.S. adults have uploaded or shared content online including artwork, photos, videos or stories. About 18% post comments too. Overall, 31% of adults are elite tech users who are heavy and frequent internet users that are also engaged with user-generated content.

Web publishers of all sizes are expanding beyond their own content and services. We are witnessing an arms race where publishers compete by installing various user-generated content sharing features in an attempt to increase page views, time spent, and to grow revenues in the expanding marketplace.

Add all this user-contributed content together with Web publisher content, and you have information overload that must be managed more effectively. It's safe to say that suppliers of social systems will be kept very busy. To succeed, they are developing and delivering easily-implemented services on sites. There are clear opportunities related to site search, whether across the publisher's own content or their new social networks. Site content discovery and recommendation tools are also getting delivered to publishers.

Consumers are the ultimate beneficiaries in this early race to include their feedback into search algorithms. The technologies that prevail – whether explicit, implicit or a combination – will be the ones that enable consumers see their intentions reflected in their search results, not just their keywords.