



VIGNETTE[®]
*Bringing content to **life***

In Search of the Human Element

Why Social Search will Solve the Enterprise Search Challenge

WHITE PAPER

“Winston fell asleep murmuring ‘Sanity is not statistical,’ with the feeling that this remark contained in it a profound wisdom.”

– George Orwell excerpt from the novel “1984”

While George Orwell didn’t specifically mention enterprise search in his famous vision of the future (written in 1949), he still managed to convey a truth that resonates today: When it comes to relevant search query results, “it’s not about the statistics.” Sanity comes from the human element.

Until recently, this concept was mostly ignored by enterprise search solutions. Instead, search was based on text-matching algorithms and models that methodically sifted through link structures or categorization schemes. After analysis and compilation, a surge of search results (all quantifiably correct) were released. While a statistician might be impressed by the outcome, the user who initiated the search simply drowned in a sea of homogenous data.

The question became, “Is it possible to produce a set of meaningful search results that will help a user-base, rather than inundate them?”

In response, enterprise search technologies introduced tools that allowed “experts” to tune the search algorithm for their specific content set, and introduced meta-tagging best practices to structure content so that more meaningful set of results would return. More recently, explicit user actions such as click-throughs, ratings, and feedback were introduced to solve the issue of search relevancy. Today, advanced techniques including “social search” have evolved to take into account how humans search for, find, and consume information and products in the physical world.

This white paper examines the ever-evolving world of enterprise search, and introduces a unique approach called Social Search that mines a collective set of user behavior patterns to produce search results directly pertinent to the needs of an individual user.

Comparing Various Approaches to Solve Search

Enterprise Search has long been the de facto standard for corporate knowledge databases and Web sites. Because this approach is typically based on text-matching algorithms, it does a thorough job in finding all documents that match a particular query term. Some highly evolved search engines use specialized classification systems or pattern recognition techniques that rely on statistical inference. As a result, enterprise search engines have experienced huge increases in performance, comprehensiveness, and automation. However, they still lack the single most important ingredient that produces relevant search results: subjectivity. Based upon a few ambiguous words typed into a text field, a search engine still has no reliable way of accurately interpreting the actual intent of an individual user.

To cope with subjectivity, and the lack of understanding around user intent, various approaches have been developed to try and provide more useful set of results to the end user:

Expert-based Tuning

Tuning the Algorithms. Enterprise search technologies introduced tools that allowed search administrators and content architects to select the specific algorithms that the search engine would use (e.g. Bayesian, Neural Network, Natural Language) and in what proportion they would be used.

These tools further allowed administrators to decide how much relative weighting the algorithms should give to a document’s title, description, body text, and document attributes (e.g. size, date, format) in determining whether a particular document matches the user’s query.

These tools also allowed administrators to manually force certain documents or whole collections of content to the top of results given the user's query—this was frequently used as an "override" when the search engine would not return content that authors or Web site administrators intuitively felt was the best match.

The reliance on experts and tools to tune search results quickly became a sore point for most businesses. Search administrators were rarely experts in the content's subject matter, nor did they possess any special knowledge about how visitors thought about their company's products and services which made it difficult to establish rules that helped users. Also, because the content and queries are constantly changing and evolving on the Web site, manually tuning the system simply was not a scalable strategy.

Tuning the Content. Search technologists also believed that if content could be better structured before it was indexed by the engine that better results would flow out—otherwise the mantra of "garbage in, garbage out" would continue to be true. Organizations therefore implemented meta-tagging and classification best practices.

Meta-tagging and classification is the process of applying identifying labels, category codes, titles, and descriptions (sometimes called metadata) to documents. As a result, the content will, in theory, be more accurately recognized by a search engine. This process can take advantage of proven, existing knowledge or expertise possessed by internal experts, that isn't possessed by the average Web site user.

This technique can also be a manually intensive process that requires dedicated experts or librarians to tag and maintain the content. For large dynamic Web sites, this is an overwhelming task. In addition, Web site authors and taxonomists intentionally and unintentionally bias the metadata to reflect a particular perspective or preference that might adversely affect the relevancy of the documents to the average user.

To solve problems of scale and bias, automated meta-tagging technologies and processes were established. However this approach frequently leads to many of the same formula-based relevancy and interpretation issues encountered by standard search engines as previously described.

Tuning the Queries. Enterprise search technologists also believed that if they could shape a user's query to better reflect what content is available, then they could do a better job at answering the question the user should have asked. Stemming, tokenization, segmentation, part-of-speech recognition and tagging, query expansion, and a myriad of other techniques are used to preprocess queries before they are sent to the search engine. More fundamentally, most search engines have thesaurus capabilities which allow a single query to act as several queries at once. For example, a user's query for telephone might also be expanded to include cell phone, mobile phone, long distance, Motorola, Nokia, AT&T, and so on.

Query tuning ultimately seeks to derive a common language with the visitor, but unfortunately these tuning techniques are only providing semantic expansion of the matching possibilities rather than genuinely understanding the intent of the visitor and pairing them with the proper content. For example, if users who searched for "copper pipes" ultimately want information on how to repair a broken sink, then query tuning is still not deciphering the true intent of the user.

User-explicit Tuning

Expert-based approaches proved to be too rigid and laborious over time and so a more scalable model was envisioned; one that would take into account user feedback. The hope was that search results could be improved by incorporating user judgment into the search process such as tracking clicks on search results or asking for explicit user ratings and comments about the quality of results. User-based models gained attention because they focus on the "human element": incorporating behaviors of Web site visitors.

User-centric search applications observe and analyze explicit visitor actions such as "clicks-throughs," voting, and visitor comments. More recently, some engines have begun incorporating social tagging. Made popular by del.icio.us and other Web 2.0 Web sites in the mid-2000s, social tagging is a decentralized way to tag Web site content. Fundamentally content is still manually organized by explicit human judgment, but in this case, it is the end users are tagging content instead of internal experts.

By monitoring this feedback and behavior, user-centric models can tap into a deep well of human experience that reflects a wide range of subjective opinions and judgment related to the usefulness of the selected material. Also, because the feedback is dynamic, search results are more current than static rules or metadata.

There are still drawbacks, however:

Broken feedback loops. Because most user-centric search systems monitor only a limited set of explicit user actions, there is a large potential for error. Because “click-throughs” are erroneously considered a successful user query, search systems will give credit to that bad result which will cause it to appear higher in the search results over time, and making it a more popular document in general.

This popularity also creates a secondary hazard with tagging. In effect, the more popular the content is, the more diverse the set of subjects users tag it with, and the more likely it will show up in any given search result. In both cases, “most popular” doesn’t necessarily equate to “most relevant.”

Survey bias. Social Search results that require explicit user feedback can also be skewed because they may only take into account the behavior of a very small sample of the overall Web site visitor population. Deemed the 1:99 problem, systems that require explicit user feedback tend to attract the 1% of users who are most polarized and therefore are a poor representation of your user base, while ignoring the 99% of visitors who represent the silent majority.

Perhaps most concerning is that user-centric systems that rely on explicit feedback have severe exposure to gaming. Malicious users can manipulate the behavior of the live site using automated link bots, form bots, tagging engines that impersonate users, and various other techniques.

Social Search and the “Wisdom of Crowds”

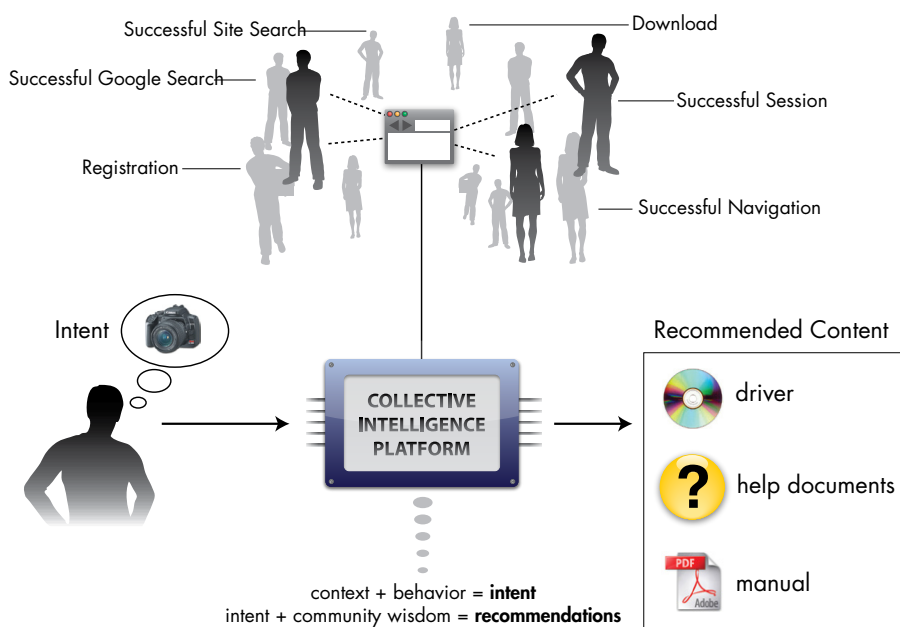
In 2004, journalist James Surowiecki wrote a book called “The Wisdom of the Crowds,” in which he proposed that small groups are often more intelligent than isolated individuals. Surowiecki said that there are four essential elements that form a “wise crowd”: diversity of opinion, independence, decentralization, and aggregation. His theory highlights the importance of understanding of how and where information is obtained: preferably from a diverse, decentralized crowd sampling.

A search engine that takes into account various forms of user behavior, while respecting Surowiecki’s concepts, is what we will now define as Social Search. Most social search techniques rely on explicit (and often misleading) user actions and feedback however. Social search techniques also frequently ignore a key element in determining a visitor’s true intent when conducting a search: the overall context of a visitor’s behavior. Context, in the case of social search, refers not only to the context of the visitor’s individual actions on Web site, but also how those actions compare and relate to the actions of other visitors.

In addition to the aforementioned criteria, there is one other important ingredient that can benefit social search: silent observation. This approach contrasts significantly with most existing social search monitoring techniques in which behavior is collected via explicit user actions (such as tagging, voting, feedback, etc.) The goal of silent observation, however, is to accurately interpret and determine the true intent of a Web site visitor without interrupting the flow of the experience. This can be accomplished by observing certain key implicit visitor behaviors such as:

- How a visitor arrives at the site
- Registration input
- Successful and unsuccessful search queries
- Virtual bookmarks
- Navigation patterns
- Virtual prints
- Downloads
- Shopping cart actions
- Purchases

This data, along with many other types of implicit on-page and pan-page actions, can be collected from every Web site visitor and analyzed. As this data is continuously distilled, virtual communities of like-minded visitors begin to emerge. Actions, patterns, and tendencies associated with these communities form the basis of a collective perspective. This is Vignette’s philosophy and approach.



This approach can result in a magnitude improvement in conversion rates versus standard search engines. The advantages to this approach have not gone unnoticed in businesses.

Social science experts have observed that every person plays multiple roles in life and the context of our current role drives our needs and desires. It is with this concept in mind that Vignette has fully integrated the unique features of the “wisdom of the crowd” within its social search solution. Vignette is particularly focused on the concepts of context and intent—an approach that is missing in other search solutions.

Vignette Recommendations not only interprets behavioral observations, but also understands the context of the behavior. By combining context with behavior, Vignette Recommendations dynamically recognizes the visitor’s intent and guides them to the right content, product, or information in the Web site.

Putting Social Search to Work for You

How it Works

Vignette Recommendations combines a site’s existing search engine results with community wisdom to produce a set of optimized results that is proven to yield greater conversions, longer engagement, and improved satisfaction. Thus, Social Search can be thought of as a community layer on top of the site’s existing search engine. The original search results may be re-ordered in the process, and the augmented results may include additional results that weren’t originally produced by the search engine, but proven to be valuable to your Web site visitors. Because Vignette Recommendations is delivered as SAAS (software as a service), it can be live on a Web site in as little as 30 days with little or no development, installation or configuration.

Working inside your existing environment

Vignette Recommendations Social Search gives businesses to the full benefit of social search while keeping their current search engine in place. Social Search operates with any search engine already in place, such as Autonomy, Google, Endeca, FAST, Lucene/Nutch, and others. The service can also be configured with “platforms” including ATG, RightNow, Vignette, and other marketing and support content repositories.

	Social Search	Traditional Search Engine
Guiding Principle	UseRank™ Dynamically ranks search results based on how useful they are to site visitors. This approach is superior to keyword-based techniques because visitors are shown the content that best matches their intent, not just what keywords appear in the content.	Keyword match Matches keyword queries to actual terms and phrases within the document itself using a number of linguistic and semantic techniques. This approach will retrieve all documents that match the keyword, but not necessarily documents that are useful.
Real-time Adaptation and Tuning	Results change automatically as visitor behaviors change, continuously optimizing the conversion potential of search.	Search results change as system administrators make changes, resulting in significant time lags before problems are detected and changes are made.
User-Feedback Model	Takes into account visitor’s implicit actions throughout the Web site to optimize results	Either ignores user feedback, or requires explicit feedback leading to relevancy, bias and gaming issues
Inclusion of New Content	Automatically discovers content as soon it is viewed by visitors	Requires re-spidering of the Web site a periodic basis, or integration into content publishing system
Federation	Automatically federates search results across multiple content stores, Web site, and event different search engines based on usage. No federation “broker” required.	Requires a search broker that manages the distribution of queries, and the assimilation and ranking of results. Requires customized integration into most content repositories
Inclusion of Browsing Behavior	Goes beyond simple click-throughs, taking into account successful and unsuccessful browsing behavior throughout the site to continuously improve search	Ignores browsing behavior
Integration into Existing Environment	Works with any existing search engine; no change to internal systems	Requires the old search engine be uninstalled, and the new engine to be installed, configured, and tuned.
Support for All Asset Types	Yes	Partial
Support for All Languages	Yes	Partial

Conclusion

The exponential growth of content and dynamic nature of today's Web sites requires a fundamental rethinking of how to determine the intent of Web site users and match them with the most appropriate content. Enterprise search technologists have begun asking these questions, but the industry is still early in its transition from an expert-centric to a user-centric mentality. The ultimate solution will require that search systems be highly scalable, intelligent, and free of bias.

Social Search is by definition labor-scalable, content-scalable, and highly-dynamic. Social Search also benefits from the collective intelligence of site visitors, creating a more natural and compelling search experience than what is possible through experts only. Social Search, if conceived of properly, will also be devoid of bias by utilizing the silent behaviors of Web site users which reflect which content truly satisfies which intent, while preventing gaming tactics or over representation of the vocal minority.