Market Niche Discovery with Social Network Analysis Techniques

Dale Hunscher
SI 614 W06
April 23, 2006

Introduction
Social bookmarking sites such as http://del.icio.us/ (Delicious) and http://www.technorati.com (Technorati) exemplify a vaguely defined suite of technologies and applications known as Web 2.0, which implies that they are both collaborative by nature and imbued with semantic enhancements. Bookmarking sites permit site members to store online the sort of bookmarks normally kept in their Web browser, making them available from any computer on which the member accesses the Internet. In addition to making the bookmarks ubiquitous, bookmarking sites allow the member to attach both a description and one or more textual tokens known as tags. The social element of these sites arises from the fact that all members’ bookmarks are visible to the public, and are findable either by searching on a tag employed by the member or by accessing a URL consisting of the site’s domain address plus the member’s name. For example, the author’s Delicious bookmarks are accessible at http://del.icio.us/DHunscher/

Bookmarking sites also allow the member to enter a descriptive profile that can be as revealing or as opaque as one chooses. Contact information can be made available in this fashion, allowing members with similar interests to communicate directly with each other. It is also possible to indicate interest in the bookmarks associated either with a particular tag or another member with similar interests, and have sites tagged with the tag or by the member appear in one’s Inbox.

Tags provide a crude but useful classification scheme, often referred to (albeit inaccurately) as folk taxonomies or folksonomies. Unlike true taxonomies, tagging schemes are non-hierarchical, possessing no features for the representation of either synonymy, hypernymy/hyponymy (genus/species), or holonymy/meronymy (whole/part) relationships. In fact each tag stands alone, with no syntactic or semantic relationship of any kind defined between it and other tags. The bookmarking sites provide mechanisms whereby one can search for URLs by tag to find all bookmarks with that tag, or to search by URL and find all bookmarks of that URL. In either case, the result set allows you to know not only who else is using a tag or bookmarking a URL, but how others describe the URL and what other tags they have used to mark it. Bookmarking sites make it easy to copy URL entries, including the tags employed by the originating member.

Collectively, these mechanisms simultaneously facilitate discovery of URLs of personal interest, other members with similar interests, other tags used in conjunction with a given tag, sets of URLs tagged with a given tag, or other tags used by the community as search markers for a given site. Features are provided in the result sets that facilitate search continuation along any of these dimensions, it is possible to engage in exploratory search, shifting from one search facet to another at will, consistent with the berry-picking metaphor for online search set forth by Bates (1989).

A number of recent papers have explored the possibilities and limitations inherent in bookmarking sites, with special emphasis on Delicious as the oldest and most mature example, e.g., Golder & Huberman (2006). Many of the insights taken from these works have to do with the pivotal role of tags. Tags mean many different things with respect to the URLs to which they are applied. Golder and Huberman identified seven tasks tags can perform for bookmarks: associating topics with the material referenced by the URL; identifying for the bookmark;
identifying the owner of the material; refining or qualifying the meaning of other tags applied to the material; identifying qualities or characteristics of the material; asserting ownership of the material by the tagger him- or herself; and organization of tasks performed with or on the material. As in Golder and Huberman’s research, this analysis views tags primarily from a structural perspective, focusing on objective factors such as indegree and edge weight, but also attempts to relate these to the semantics of the tags.

**Data Set Description**

Bookmarking data are structured as a tripartite network consisting of directed Agent-Concept-Entity relationships. Agents mentally assign concepts to an entity to categorize it or associate some characteristic with it. An entity can be anything referred to by a noun, e.g., an object, another agent, an event, or an abstract concept. In our case, a site registrant or member (sometimes referred to herein as a *name* when referencing the equivalent entity in the data set) bookmarks a Universal Resource Locator or *URL*, associating concepts with it via tokens known as *tags*. The relationships between the entities are many-to-many: many members can use the same tag, and each member can use many tags; likewise, a tag may be assigned to many sites, and a site may have many different tags applied to it.

**Data Acquisition Methodology**

Delicious search pages for the tag *nokia770* were manually downloaded to act the starting point for the data set. Nokia770 refers to a Linux-based PDA that had been first announced in May 2005 and on the market since November 2005 in Europe and since early January worldwide. As of March 6, 2006, when the search was conducted, there were 517 Delicious entries tagged with the nokia770 tag. In many cases, multiple entries referred to the same URL by different titles, since members can enter a title of their choice, but each member choosing to bookmark the URL could enter whatever title, description, and tags they chose. As the collective set of tags associated with a URL grows, Delicious begins to suggest the most popular tags, leading to a rich-get-richer phenomenon.

A Perl program was written that parsed the search pages, found the hyperlink address of the Delicious page for the URL, and for each such URL retrieved the Delicious pages that showed all members linking to the URL. Another Perl script was used to locate the entries for all members who applied nokia770 as a tag for the URL. A third script parsed these references into triplets with unique URL-name-tag\(^1\) combinations derived from the original entries. Another script transformed the triplets into a Pajek project file with eight networks and four partitions:

<table>
<thead>
<tr>
<th>Program-Generated Networks</th>
<th>Partitions</th>
<th>1-Mode Networks Created [2-Mode] Networks</th>
<th>Manually Created 1-Mode Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network directed name-tag-url triplets</td>
<td>name-url-tag</td>
<td>1-Mode Networks from 2-Mode Networks</td>
<td>Created manually from 2-Mode Networks</td>
</tr>
<tr>
<td>Network undirected name-tag-url triplets</td>
<td>name-url-tag</td>
<td>Network names from name-tag</td>
<td>Network tags from name-tag</td>
</tr>
<tr>
<td>Network directed name-to-tag</td>
<td>name-tag</td>
<td>names from name-tag</td>
<td>Network tags from name-tag</td>
</tr>
<tr>
<td>Network 2-mode undirected name-to-tag [2-Mode]</td>
<td>name-tag</td>
<td>tags from name-tag</td>
<td>Network names from name-url</td>
</tr>
<tr>
<td>Network directed name-to-url</td>
<td>name-url</td>
<td>names from name-url</td>
<td>Network urls from name-url</td>
</tr>
<tr>
<td>Network 2-mode undirected name-to-url [2-Mode]</td>
<td>name-url</td>
<td>urls from name-url</td>
<td>Network tags from tag-url</td>
</tr>
<tr>
<td>Network directed tag-to-url</td>
<td>tag-url</td>
<td>tags from tag-url</td>
<td>Network urls from tag-url</td>
</tr>
<tr>
<td>Network 2-mode undirected tag-to-url [2-Mode]</td>
<td>tag-url</td>
<td>urls from tag-url</td>
<td>Networks from tag-url</td>
</tr>
</tbody>
</table>

Multiple arcs between nodes were not permitted, e.g., those representing a member applying the same tag to multiple URLs. Instead, the cardinality of multiple arcs became the weight of a

---

\(^1\) The term *name* was used to refer to the member’s “handle”, a text token used to identify the member within the Delicious site.
single arc. Pajek’s facilities were then employed to create manually the six 1-mode networks that can be derived from the program-generated 2-mode networks.

These networks contained 233 URLs, 139 member names (anonymized by the Perl scripts), and 684 unique tags. These networks and partitions formed the data set used in the analytic phase.

**Analysis**

**Distributions within Networks in the Data Set**

Correlations of the various entity relations produced a few puzzles and some new insights. The all-degree frequency of this network is a classic power law distribution. This is the result of a “rich get richer” phenomenon caused by several aspects of Delicious’ behavior. First, the tags used in association with the earliest URLs to appear with information about the Nokia 770 were at an advantage in that Delicious begins to suggest tags for the site, making the earliest tags the easiest to apply. Second, a member can query on a likely tag to see who else has used the tag, on what URLs they chose to apply the tag, and what other tags they applied, causing URLs other members associated with the tag to float to the surface, as it were, of the vast pool of sites on the Web. Third, once having discovered URLs via such a query, the member can easily copy the entry of another member and apply the tags that member employed to tag the site.

There are three one-mode relationships (name-tag, tag-url, and name-url) that can be derived from the 2-mode networks created from the data set. Thus for each entity, there are two relationships in which it participates, and two data series representing those relationships. The next step was to look at the correlation between an entity’s position within each of the two series in which it takes part.

**URLs**

The indegree of a URL from a given tag shows the frequency with which members associated the concept represented by the tag with the URL. The indegree of the URL from a given name shows the number of members who chose to bookmark the URL.

This chart and the next two show the two series under consideration with each series sorted by rank. Of the 230+ URLs, just over 200 had an indegree from tags in the range 195-230. From this rank onward, the degree falls off steeply. This appears to indicate that the vast majority of URLs are tagged at consistent levels across the member community. One thought about the steep drop-off is that the URLs on the side of the “cliff” may be much more recent than the others, but there was not enough time to go back and correlate age with indegree.
The programmatically derived name-to-url bipartite network shows a weak power law curve for URL indegree, similar to that of the indegree of tags referenced by members. The correlation between the two URL data series was weak but still positive.

Names
The outdegree of names pointing to URLs is indicative of the member’s breadth of interest in the domain represented the URLs in the data set. The outdegree of names to tags, on the other hand, signifies the member’s perception of the conceptual richness of domain.

The relation between outdegree of members referencing a URL and their ranks was inverse and nearly linear. This suggests that almost everyone who used Delicious to tag Nokia 770-related sites tagged multiple such sites; there is no long tail of marginally interested members. The same applies to names referencing tags, but even more so; of about 140 members, over 130 used 50 or more different tags on their nokia770-related sites.

Tags
Tags mediate the relationship between members and URLs in the tripartite network. The indegree of a tag represents the overall popularity of the concept represented by the tag, and the outdegree corresponds to the perceived applicability of the concept the tag signifies to the domain represented by the URLs in the data set.

The indegree of tags referenced by name presents a shallow nonlinear curve, while the outdegree of tags referencing URLs shows a curious set of plateaus in the midrange. This plateau effect may be the result of above-described rich-get-richer phenomenon. Tags with a higher degree than the plateau would likely represent concepts that were strongly associated with the Nokia 770 and that therefore were applied most frequently to newly discovered sites. The long tail would then represent tags that are associated with singular interests of particular members.

Analysis through Visualizations
The next step in the analysis was to attempt to visualize the name-tag-url network. In Pajek, The network was first drawn with its program-generated vertex positions as-is, then the Layout-Circular-Original command was used to create a uniform starting point for energization. Fruchterman-Reingold (F-R) energization was applied 12 times; the result is shown in Figure 1.
In this figure, vertices are color-coded pink for URLs, green for members, and blue for tags. The horizontal elongation suggested the possibility that the network was not uniform and might contain structure under the surface.

For comparison, an Erdos-Renyi random graph of similar proportions was generated and then energized in the same fashion as the name-tag-url network. Its appearance was more uniform, taking on the uniform elliptical shape shown in Figure 2.

An all-degree partition and its associated vector were created, and the network was redrawn using the original name-tag-url partition coupled with the all-degree vector, which produced the result shown in Figure 3 (no further energization was applied).
Vertices are color-coded as in Figure 1. From this it was possible to see that nokia770 was the most-used tag (no great surprise), and user3307 was the member who had applied the most tags. The site which had been tagged most extensively was unexpected: www.gutenberg.org, a source of free electronic books (also known as e-books). Given the pre-release status of the Nokia 770 through most of the data accrual period, almost all tagging activity was expected to be done by software developers and “gadget-heads”. The Nokia 770 was first announced at a Linux conference on May 25, 2005, but the developer community had come into existence informally several months prior, seeded by Nokia’s communications to its developer email lists and newsgroups.

E-books are a logical application for the 770 because it has a high-resolution 800x240 screen and still fits in one’s pocket, the best compromise yet between portability and readability from an e-book perspective. Moreover, the 770’s ubiquitous Internet access would make finding and downloading new e-books very convenient. Nonetheless, e-books are a niche market, very distinct from the application interests of the developer community. It seemed possible that there could be two distinct communities lurking under the surface of the tangled mass thus far visualized—those interested in literary applications (bookworms) and a community of members interested in development and other technical aspects of the 770 (geeks).²

Investigation of this possibility began with the name-tag network. Like the name-tag-url network, it too began to hint at underlying community. To clarify this, edges with a value of less than 3 were removed to eliminate names of members who had seldom employed a tag, and also

² The term “community” might seem too strong, since there is nothing in the data that suggests that the members of each group were aware of each other at all, much less in communication with each other, as the term implies. In many cases interest clusters might form in a bookmarking database like that underlying Delicious without the emergence of a community. In this instance, the author knows from personal experience as a member of the geek community that many of its members were in fact aware of and in communication with others. Cursory investigation showed that this also pertained to the core bookworm group members.

Beyond these circumstantial justifications for the applicability of the term, other justifications can be cited. Bookmarking sites were deliberately created to foster awareness of and communication with other members with whom one has shared interests, and there is widespread evidence that bookmarking sites are in fact used for the purpose of finding others of like mind. One example would be the common practice of bloggers putting their personal Delicious “tag clouds” on the margin of their blogs.
the URLs to which a given tag was infrequently applied. An all-degree partition was created, and now-isolated nodes removed. Upon energization, it became apparent there were indeed two communities, as shown in Figure 4.

The next task was to identify the members and tags that bridge the two networks. The original name-tag network was reduced by creating an all-degree partition and removing nodes with low degree. Experimentation revealed the bridge nodes were best isolated by reduction on degree less than 21. This network is shown in Figure 5 with the all-degree partition and normalized all-degree vector applied and energized with the Kamada-Kawai algorithm).
To assess the importance of the few nodes that connected the two communities, a betweenness vector was created, which when energized produced the visualization in Figure 6:

Three members and two tags form the bridge between the two communities. user3307 clearly plays a central role as the gatekeeper of the bookworm community. user204 appears to be a geek rather than a bookworm, though the figure contains only limited evidence to this effect—this member only references the online, opensource and reference tags in this network, suggesting that he might be looking at the 770 as a tool for accessing online references to Open Source tools, frameworks, and/or applications. user2536 is clearly in the developer community, referencing the tags opensource and linux.

**Leveraging the Significance of Tags**

It was then decided to focus on tags as a means to further insight into the two communities. Tags are indicators of the conceptual framework in which the member is operating, at least with respect to the domain under consideration. Viewed from the member’s perspective, tags denote the concepts that come to mind most often when the member deliberately seeks or serendipitously locates a new source of information. From the standpoint of a URL, tags represent the concepts members associate with the URL as a source of information. Each instance of the association of a tag with a URL is in effect a member’s vote for the salience of the concept represented by the tag.
To leverage the conceptual framework emerging from this dynamic, two facets of the tripartite network were examined: the tags each member chose most often and the tags most frequently associated with a particular URL. Both metrics are reflected in the values assigned to edges. These would provide an indication of the concepts in which members were most interested and the sites most often tagged with a particular concept. Application of Pajek’s valued core algorithm revealed members who applied the most tags to URLs also tagged by other members; this would indicate the URLs of highest interest, the underlying concepts leading to the popularity of these URLs, and the members most involved in the creation of this emergent conceptual framework.

The 35-slice of the name-tag-url was limited to roughly the top 5% of the distribution—the so-called “long tail” of the distribution graphed such that the X-axis represents the sum of values of a node’s edges and the Y-axis indicates the frequency of occurrence of a particular value sum. Energization and a bit of manual manipulation produced the visualization in Figure 7. Application of a weighted indegree vector shows a similar pattern (Figure 8).
Viewing the context of the three free e-book sites, gutenberg.org, memoware.com, and manybooks.net in this manner showed that the popularity of the sites and the diversity within the set of shared concepts members applied frequently to the sites are the social forces creating the bookworm community. The tags pda, reference, books, literature, ebook, and ebooks were used very frequently to tag these sites. At this m-slice level, members have disappeared entirely from the bookworm community, indicating that the bookworms tag at a lower frequency level than at least some of their counterparts among the geeks, where eight members still appear in this m-slice. This suggests that the shared conceptual framework of the bookworms has greater internal consistency and intensity of focus than that of the geeks.

Community Emergence

From the analysis thus far it was clear that two communities, bookworms and geeks, did exist. The existence of the bookworm community had been a surprise in the first place, so it seemed appropriate to hypothesize that it was a late-emerging phenomenon. This hypothesis could be tested visually by looking at the data as it accrued over time, so the Perl script used to create the Pajek project file was enhanced to include time series data. The Delicious URL pages listed tagging of a URL by month, hence this was the time unit employed as the “clock tick” for the time series. This data enabled generation of twenty-nine networks that showed the emergence of the two communities over time.

The resulting visualizations disproved the hypothesis stated above. The bookworm community was already

---

3 A fourth e-book site, http://only.mawhrin.net/fbreader/, does not appear at this level of resolution, but was also the target of extensive interest by the bookwork community. FBReader is an e-book reader exclusively targeted at the Linux platform and is widely considered very suitable for use on the 770 device.
strong well in advance of the release of the Nokia 770, with Delicious bookmarking preceding its announcement by almost two years. These visualizations began with the original name-tag-url network; edges < 5 were removed along with isolated nodes.

In Figure 9, all three of the most popular bookworm sites are already present, connected to a few geek nodes by what will become the most popular bridge concepts. One of the earliest geek URLs, scratchbox.org, is an Open Source Linux cross-compilation environment that would soon become the basis for development of the Nokia 770 software development effort.

Eight months later, the bookworm community is in full flower, while the geeks are growing slowly, as shown in Figure 10. Figure 11 shows the network at the inception of public interest in the new device. Maemo.org, the Open Source effort creating Nokia 770 operating system and graphical user interface, has appeared on the geek side. The bookworms have proliferated links as well, but their growth pace is slowing just as the geeks start gaining momentum. It is no coincidence that the nokia770 tag first appears in this time slot.

Figure 12 shows the point in time where the 770 is publicly available worldwide. The bookworm community has changed very little, and thanks to its rapid growth, the geek community is now predominant.

**Speculations**

It was observed earlier that the shared conceptual framework of the bookworms, their community *Weltanschauung*, showed greater internal consistency and intensity of focus than that of the geeks. This raises some interesting questions, full-fledged answers to which are outside the scope of this paper, but some speculation is appropriate at this point. By whom, in what way, and to what end could such knowledge be applied? One possibility comes to mind. In the circumstance examined herein, the manufacturer of this new class of product could have applied social network analysis to the wealth of available bookmark tagging and blog categorization data and observed the spontaneous emergence of the two communities.
The bookworm community flourished prior to and independent of any public knowledge of the Nokia 770 product. The emergence of the 770 generated strong interest within the segment of the bookworm community represented in our sample, but the 770 does not appear to have stimulated further growth of any significance in that community. The bookworms' idea of a “killer app”, the delivery of electronic books, could be the Nokia 770’s leg up on entrenched competitive devices such as the various Pocket PC and Palm Pilot models. In the most optimistic case, given the popularity of audio books and the continued growth of the hard-copy book market, the 770 could conceivably ride the wave of an explosively expanding new market. At minimum, the bookworms have great potential to aid in the formation of a critical mass of devices in the marketplace that would stimulate development of additional applications leveraging the 770’s unique characteristics, applications that could take the 770 into the mainstream in effective competition with the entrenched PDA brands.

Unfortunately, Nokia has either not gained the insights found in this analysis or has chosen not to act on them. Failure to publicize and exploit the e-book application may be leaving the Nokia 770 Internet Tablet in the unenviable role of a technology in search of a problem to solve.

**Next Steps**

The focus of this paper shifted during the execution, so only a cursory literature search has been performed in the area of application of network theory to marketing, but even a cursory search shows there has been much work and there still is much interest in this area. Social network analysis has been on the radar of market researchers since at least the 1980’s, for example Carroll et al. (1986); Hoffman & Frank (1986); Hutt & Reingen (1987); and Hutt et al. (1988). More recent research in this area has focused on the application of disease contagion models to such social software mechanisms as recommender systems, e.g., Leskovec et al. (2005).

Leskovec and his colleagues found that recommender systems had limited effect, in that more recommendations led to more purchases only to a certain point, after which additional recommendations led to little or no additional likelihood of purchase by the recommendation recipient. Their study focused on actual purchasers of products, looking only at their recommending behavior and the purchasing behavior of the recipients of their recommendations, without factoring in the normative characteristics of senders, recipients, or products in a manner analogous to tagging behavior. By comparison, the approach taken in this investigation is looking in the marketplace for the spontaneous and proactive involvement of unanticipated groups of prospective purchasers, the encouragement of which could provide a previously unexpected source of sales and recommendations.

The analytical process described herein was exploratory and largely aimed at acquiring mastery of the tools and techniques of social network analysis. However, with further study of the literature and both retrospective and prospective research, it is possible that a more rigorous conceptual model and methodology building on the foundations laid by Leskovec et al. and others that would enable systematic evaluation of similar market opportunities using Web-based social networking data such as that provided by Delicious and Technorati, not to mention the rich lode of information that can be derived from other social software data sources such as blogs and wikis.

---

4 The author is choosing to ignore the possibility that the insights themselves could be erroneous. 😊
References


