

Unraveling the Taste Fabric of Social Networks

Hugo Liu, Pattie Maes, Glorianna Davenport

The Media Laboratory, Massachusetts Institute of Technology

e-mail: {hugo, pattie, gid}@media.mit.edu

Abstract

Popular online social networks such as Friendster and MySpace do more than simply reveal the superficial structure of social connectedness—the rich meanings bottled within social network profiles themselves imply deeper patterns of culture and taste. If these latent semantic fabrics of taste could be harvested formally, the resultant resource would afford completely novel ways for representing and reasoning about web users and people in general. This paper narrates the theory and technique of such a feat—the natural language text of 100,000 social network profiles were captured, mapped into a diverse ontology of music, books, films, foods, etc., and machine learning was applied to infer a semantic fabric of taste. Taste fabrics bring us closer to improvisational manipulations of meaning, and afford us at least three semantic functions—the creation of semantically flexible user representations, cross-domain taste-based recommendation, and the computation of taste-similarity between people—whose use cases are demonstrated within the context of three applications—the InterestMap, Ambient Semantics, and IdentityMirror. Finally, we evaluate the quality of the taste fabrics, and distill from this research reusable methodologies and techniques of consequence to the semantic mining and Semantic Web communities.

Keywords

Social Networks, Semantic Mediation, Culture and Taste, Ethotic Representation, Recommender Systems, Latent Semantics, User Modeling, Relational Mining, Computational Aesthetics, Psychographics.

1 Introduction

Recently, an online social network phenomenon has swept over the Web—MySpace, Friendster, Orkut, thefacebook, LinkedIn—and the signs say that social networks are here to stay; they constitute the *social Semantic Web*. Few could have imagined it—tens of millions of Web users joining these social network sites, listing openly their online friends and enlisting offline ones too, and more often than not, specifying in great detail and with apparent exhibitionism tidbits about who they are, what music they listen to, what films they fancy. Erstwhile, computer scientists were struggling to extract user profiles by scraping personal homepages, but now, the extraction task is greatly simplified. Not only do self-described personal social network profiles avail greater detail about a user's interests than a homepage, but on the three most popular sites, these interests are distributed across a greater spectrum of interests such as books, music, films, television shows, foods, sports, passions, profession, etc. Furthermore, the presentation of these user interests are greatly condensed. Whereas interests are sprinkled across hard-to-parse natural language text on personal homepages, the prevailing convention on social network profiles sees interests given as punctuation-delimited keywords and keyphrases (see examples of profiles in Figure 1), sorted by interest genres.

It could be argued that online social networks reflect—with a great degree of insight—the social and cultural order of offline society in general, though we readily concede that not all social segments are fairly represented. Notwithstanding, social network profiles are still a goldmine of information about people and socialization. Much computational research has aimed to understand and model the surface connectedness and social clustering of people within online social network through the application of graph theory to friend-relationships (Wasserman, 1994; Jensen & Neville, 2002; McCallum, Corrada-Emmanuel & Wang, 2005); ethnographers are finding these networks new resources for studying social behavior in-the-wild. Online social networks have also implemented site features that allow persons to be searched or matched with others on the basis of shared interest keywords.

Liminal semantics. However, the full depth of the semantics contained within social network profiles has been under-explored. This paper narrates one such deep semantic exploration of social network profiles. Under the keyword mediation scheme, a person who likes “rock climbing” will miss the opportunity to be connected with a friend-of-a-friend (foaf) who likes “wakeboarding” because keyword-based search is vulnerable to the *semantic gap* problem. We envision that persons who like “rock climbing” and “wakeboarding” should be matched on the basis of them both enjoying common *ethoi* (characteristics) such as “sense of adventure,” “outdoor sports,” “and “thrill seeking.” A critic might at this point suggest that this could all be achieved through the semantic mediation of an organizing ontology in which both “rock climbing” and “wakeboarding” are subordinate to the common governor, “outdoor sports.” While we agree that *a priori* ontologies can mediate, and in fact they play a part in this paper’s research, there are subtler examples where *a priori* ontologies would always fail. For example, consider that “rock climbing,” “yoga,” the food “sushi,” the music of “Mozart,” and the books of “Ralph Waldo Emerson” all have something in common. But we cannot expect *a priori* ontologies to anticipate such ephemeral affinities between these items. The common threads that weave these items have the qualities of being liminal (barely perceptible), affective (emotional), and exhibit shared identity, culture, and taste. In short, these items are held together by a liminal semantic force field, and united they constitute a *taste ethos*.

What is a taste ethos? A taste ethos is an ephemeral clustering of interests from the taste fabric. Later in this paper we will formally explain and justify inferring a taste fabric from social network profiles, but for now, it suffices to say that the taste fabric is an n by n correlation matrix, for all n interest items mentioned or implied on a social network (e.g. a book title, a book author, a musician, a type of food, etc.). Taste fabric specifies the pairwise affinity between *any* two interest items, using a standard machine learning numeric metric known as pointwise mutual information (PMI) (Church & Hanks, 1990). If a taste fabric is an oracle which gives us the affinity between interest items as $a(x_i, x_j)$, and a taste ethos is some set of interest items x_1, x_2, \dots, x_k , then we can evaluate quantitatively the strength, or *taste-cohesiveness*, of this taste ethos. While some sets of interest items will be weakly cohesive, other sets will demonstrate strong cohesion. Using *morphological opening* and *thresholding* (Serra, 1982; Haralick, Sternberg & Zhuang, 1987), standard techniques for object recognition in the image processing field, we can discover increasingly larger sets of strong cohesiveness. The largest and most stable of these we term *taste neighborhoods*—they signify culturally stable cliques of taste. Visualizing these interconnected neighborhoods of taste, we see that it resembles a topological map of taste space!

Taste neighborhoods and taste *ethoi*, we suggest, are novel and deep mechanisms for taste-based intrapersonal and interpersonal semantic mediation. Rather than mapping two persons into interest keyword space, or into *a priori* ontological space, the approach advocated in this paper is to map the two persons first into taste-space, and then to use their shared *ethoi* and *neighborhoods* to remark about the taste-similarity of these persons.

Emergent and implicit semantics. While our work builds directly upon age-old language modeling techniques in Computational Linguistics, and graph-based associative reasoning in Artificial Intelligence (Collins & Loftus, 1975; Fellbaum, 1998; Liu & Singh, 2004), it is also sympathetic to trends in the Semantic Web literature—away from formal semantics, and toward an embrace of emergent and implicit semantics. In Volume 1 of this journal, Sheth, Ramakrishnan & Thomas (2005) distinguishes between formal, implicit, and powerful (soft) semantics for the Semantic Web movement. Whereas formal semantics must be manually specified, implicit semantics can be readily mined out of the unstructured Web using statistical approaches. Upon further refinement, implicit semantic resources can be transformed into powerful (soft) semantic resources that afford the ability to mediate informal and formal entities. Related to implicit semantics, emergent semantics is an evolutionary approach to knowledge management (Staab *et al.*, 2002; Aberer *et al.*, 2004) that advocates semantic organization to be shaped from the ground-up, *a posteriori*, and in accordance with the natural tendencies of the unstructured data—such a resource is often called a *folksonomy*. We suggest that online social network profiles give an implicit semantics for cultural taste-space, and that taste fabrics afford a semi-formal, soft semantics appropriate for semantic mediation between informal and formal entities. Finally, arising out of correlation analysis, topological features of the taste fabric—such as taste neighborhoods, identity hubs, and taste cliques—constitute an emergent semantics for taste-space.

Paper's organization. The rest of the paper has the following organization. Section Two lays out a theoretical foundation for representing and computing taste, framed within theories in the psychological and sociological literatures. In particular, it addresses a central premise of our taste-mining approach—“is the collocation of interest keywords within a single user's profile meaningful; how does that tell us anything about the fabric of taste?” Section Three narrates the computational architecture of the implementation of taste fabric, including techniques for ontology-driven natural language normalization, and taste neighborhood discovery. Section Four describes three semantic functions of a taste fabric—semantically flexible user modeling, taste-based recommendation, and interpersonal taste-similarity—within the context of three applications—InterestMap (Liu & Maes, 2005a), Ambient Semantics (Maes *et al.*, 2005), and IdentityMirror. Section Five evaluates the quality of the taste fabric by examining its efficacy in a recommendation task, and also entertains an advanced discussion apropos related work and reusable methodologies distilled from this research. The paper concludes in Section Six.

2 Theoretical Background

This section lays a theoretical foundation for how taste, identity, and social network politics are approached in this work. For the purposes of the ensuing theoretical discussion, social network profiles of concern to this project can be conceptualized as a bag of interest items which a user has written herself in natural language. In essence, it is a self-descriptive free-text user

representation, or harkening to Julie Andrews in *The Sound of Music*, “these are a few of my favorite things.” A central theoretical premise of mining taste fabric from social network profiles by discovering latent semantic correlations between interest items is that “the collocation of a user’s bag of interest items is meaningful, structured by his identity, closed within his aesthetics, and informs the total space of taste.” Section 2.1 argues that a user’s bag of interests gives a true representation of his identity, and enjoys unified ethos, or, *aesthetic closure*. Section 2.2 plays devil’s advocate and betrays some limitations to our theoretical posture. Section 2.3 theorizes a segregation of user’s profile keywords into two species—identity-level items versus interest-level items. This distinction has implications for the topological structure of the taste fabric.

2.1 Authentic identity and aesthetic closure

In the wake of this consumer-driven contemporary world, the proverb “you are what you eat” is as true as it has ever been—we are what we *consume*. Whereas there was a time in the past when people could be ontologized according to social class, psychological types, and generations—the so-called demographic categories—today’s world is filled with multiplicity, heterogeneity, and diversity. The idea that we now have a much more fine-grained vocabulary for expressing the self is what ethnographer Grant McCracken, echoing Plato, calls *plenitude* (McCracken, 1997). In a culture of plenitude, a person’s identity can only be described as the sum total of what she likes and consumes. Romantic proto-sociologist Georg Simmel (1908/1971) characterized identity using the metaphor of our life’s materials as a broken glass—in each shard, which could be our profession, our social status, our church membership, or the things we like, we see a partial reflection of our identity. The sum of these shards never fully capture our individuality, but they do begin to approach it. Simmel’s fundamental explanation of identity is Romantic in its genre. He believed that the individual, while born into the world as an unidentified *contents*, becomes over time reified into identified *forms*. Over the long run, if the individual has the opportunity to live a sufficiently diverse set of experiences (to ensure that he does not get spuriously trapped within some local maxima), the set of forms that he occupies—those shards of glass—will converge upon an authentic description of his underlying individuality. Simmel believes that the set of shards which we collect over a lifetime sum together to describe our true self because he believes in authenticity, as did Plato long before him, and Martin Heidegger after him, among others.

While Simmel postulated that earnest self-actualization would cause the collection of a person’s shards to converge upon his true individuality, the post-Freudian psychoanalyst Jacques Lacan went so far as to deny that there could be any such true individual—he carried forth the idea that the ego (self) is always constructed in the Other (culture and world’s materials). From Lacan’s work, a mediated construction theory of identity was born—the idea that who we are is wholly fabricated out of cultural materials such as language, music, books, film plots, etc. Other popular renditions of the idea that language (e.g., ontologies of music, books, etc.) controls thought include the Sapir-Whorf hypothesis, and George Orwell’s *newspeak* idea in his novel 1984. Today, mediated construction theory is carried forth primarily by the literature of feminist epistemology, but it is more or less an accepted idea.

At the end of the day, Simmel and Lacan have more in common than differences. Csikszentmihalyi and Rochberg-Halton (1981), succeed in the following reconciliation. Their

theory is that the objects that people keep in their homes, plus the things that they like and consume, constitute a “symbolic environment” which both echoes (Simmel) and reinforces (Lacan) the owner’s identity. In our work, we take a person’s social network profile to be this symbolic environment which gives a true representation of self.

If we accept that a user profile can give a true representation of self, there remains still the question of closure. Besides all being liked by a person, do the interests in his bag of interests have coherence amongst themselves? If it is the case that people tend toward a tightly unified ethos, or *aesthetic closure*, then all the interests in a person’s bag will be interconnected, interlocked, and share a common aesthetic rationale. If there is aesthetic closure, then it will be fair for our approach to regard every pair of interest co-occurrences on a profile to be significant. If we know there is not any closure, and that people are more or less arbitrary in what interests they choose, then our approach would be invalid.

Our common sense tells us that people are not completely arbitrary in what they like or consume, they hold at least partially coherent systems of opinions, personalities, ethics, and tastes, so there should be a pattern behind a person’s consumerism. The precise degree of closure, however, is proportional to at least a person’s ethicalness and perhaps his conscientiousness. In his *Ethics* (350 B.C.E.), Aristotle implied that a person’s possession of ethicalness supports closure because ethics lends a person *enkrasia* or continence, and thus the ability to be consistent. Conscientiousness, a dimension of the Big Five personality theory (John, 1990), and perhaps combined with neuroticism, a second dimension in the same theory, would lead a person to seek out consistency of judgment across his interests. They need not all fall under the genre, but they should all be of a comparable quality and enjoy a similarly high echelon of taste. Grant McCracken (1991) coined the term the Diderot Effect to describe consumers’ general compulsions for consistency—for example, John buys a new lamp that he really loves more than anything else, but when he places it in his home, he finds that his other possessions are not nearly as dear to him, so he grows unhappy with them and constantly seeks to upgrade all his possessions such that he will no longer cherish one much more than the others. Harkening to the Romantic hermeneutics of Friedrich Schleiermacher (1809/1998), we might seek to explain this compulsion for uniformity as a tendency to express a unified emotion and intention across all aspects of personhood. Indeed, McCracken himself termed this uniformity of liking the various things we consume, Diderot Unity. Diderot Unity Theory adds further support to our premise that *for the most part*, a person’s bag of interests will have aesthetic closure.

2.2 Upper bounds on theoretical ideal

From Section 2.1, we could conclude a theoretically ideal situation for our taste-mining approach—1) a user’s bag of interests is an authentic and candid representation of what the user really likes, and 2) none of the interests are out-of-place and there is strong aesthetic closure and share taste which binds together all of the interests in the bag. Here, we raise three practical problems which would degrade the theoretically ideal conditions, thus, constituting an upper bound; however, we would suggest that these would degrade but not destroy our theoretical premise, resulting in noise to be introduced into the inference of the taste fabric.

A first corruptive factor is performance. Erving Goffman (1959) poses socialization as a theatrical performance. A social network is a social setting much like Goffman’s favorite

example of a cocktail party, and in this social setting, the true self is hidden behind a number of personae or masks, where the selection of the mask to wear is constrained by the other types of people present in that setting. Goffman says that we pick our mask with the knowledge of those surrounding us, and we give a rousing performance through this mask. In other words, the socialness of the social network setting would rouse us to commit to just one of our personae, and to give a dramatic performance in line with that persona. Performance might strengthen aesthetic closure, but it could also be so overly reductive that the bag of interests no longer represent all of the aspects of the person's true identity.

A second corruptive factor is publicity. In her ethnographic review of the Friendster social networking site, Danah Boyd (2004) raises concerns over the quality and truth of profiles in light of the fact that a personal profile is public, not only to strangers, but also to one's high school friends, college friends, professors, ex-girlfriends, and coworkers alike. Because social networking sites generally make a profile visible to all these different social circles at once, Boyd suggests that some users are cowed to the fear of potentially embarrassing exposure—for example, teacher exposing to his students, or teenager exposing to his mother. As a result, users may be cowed into a lowest-common-denominator behavior, sanitizing the personal profile of all potentially embarrassing, incriminating, or offensive content.

Finally, a third corruptive factor also raised by Boyd, is the integrity and timeliness of social networks themselves. Boyd claims that Friendster profiles and friend connections are not frequently updated, leading to stale information which could distort the taste fabric if we were interested in looking at the temporal animation of the fabric. Boyd also writes about a phenomenon known as Fakesters—the creation of bogus user profiles such as for celebrities. However, the scope of Fakesters is arguably limited, and since Fakesters are chiefly imitations of actual people, aesthetic closure should still be observed and learning over Fakester profile examples should not greatly compromise the integrity of the meaning implied by the taste fabric.

2.3 Identity keywords vs. interest keywords

While each social network has an idiosyncratic representation, the common denominator across all the major web-based social networks we have examined is the representation of a person's broad interests (*e.g.* hobbies, sports, music, books, television shows, movies, and cuisines) as a set of keywords and phrases. But in addition, more than just interests, higher-level features about a person such as cultural identities (*e.g.* “raver,” “extreme sports,” “goth,” “dog lover,” “fashionista”) are also articulated via a category of *special interests* variously named, “interests,” “hobbies & interests,” or “passions.”

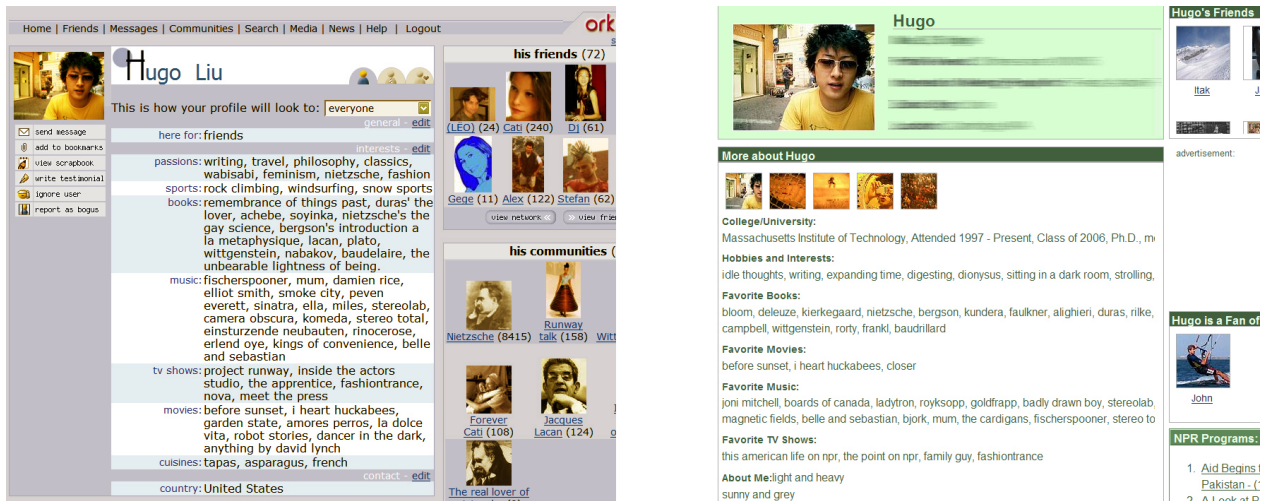


Figure 1: Examples of social network profile formats, on Orkut (*left*) and Friendster (*right*). Note the similarity of categories between the two.

As shown in the web page layout of the personal profile display (Figure 1), the *special interests* category appears above the more specific interest categories. We suggest that this placement encourages a different conceptualization for the special interests category—as a container for descriptions more central to one’s own self-concept and self-identification. Of course, syntactic and semantic requirements are not enforced regarding what can and cannot be said within any of these profile entry interfaces, but based on our experiences, with the exception of those who are intentionally tongue-and-cheek, the special interests category is usually populated with descriptors more central to the self than other categories. For example, a person may list “Nietzsche” and “Neruda” under the “books” category, and “reading,” “books,” or “literature” under the special interests category. In the normalization process of profiles, identity descriptors are inferred from descriptors listed under the special interests category (e.g. “dogs” → “Dog Lover,” “reading” → “Book Lover,” “deconstruction” → “Intellectual”).

Theoretically speaking, it is desirable to have two different granularities of description for a person. Identity descriptors are more general and constitute a far smaller ontology than interest descriptors, thus, the resulting effect is to create a taste fabric structured according to a hub-and-spoke topology. Identity descriptors serve as hubs and interest descriptors serve as spokes. The advantages to such an organization are revealed in a later section on applications, and in the evaluation of the taste fabric in a recommendation task.

Having established a theoretical premise for mining taste fabric from social network profiles, and having argued for identity descriptors to be separate from interest descriptors, the following section dives into the architecture and techniques of the taste fabric implementation.

3 Weaving the Taste Fabric

The implementation of the taste fabric making system was completed in approximately 3,000 lines of Python code. As depicted in Figure 2, the architecture for mining and weaving the taste fabric from social network profiles can be broken down into five steps: 1) acquiring the profiles

from social networking sites, 2) segmentation of the natural language profiles to produce a bag of descriptors, 3) mapping of natural language fragment descriptors into formal ontology, 4) learning the correlation matrix, and 5) discovering taste neighborhoods via morphological opening, and labeling the network topology. The following subsections examine each of these phase phases of processing more closely. A more condensed description of the mining process, sans neighborhood discovery, can be found in (Liu & Maes, 2005a).

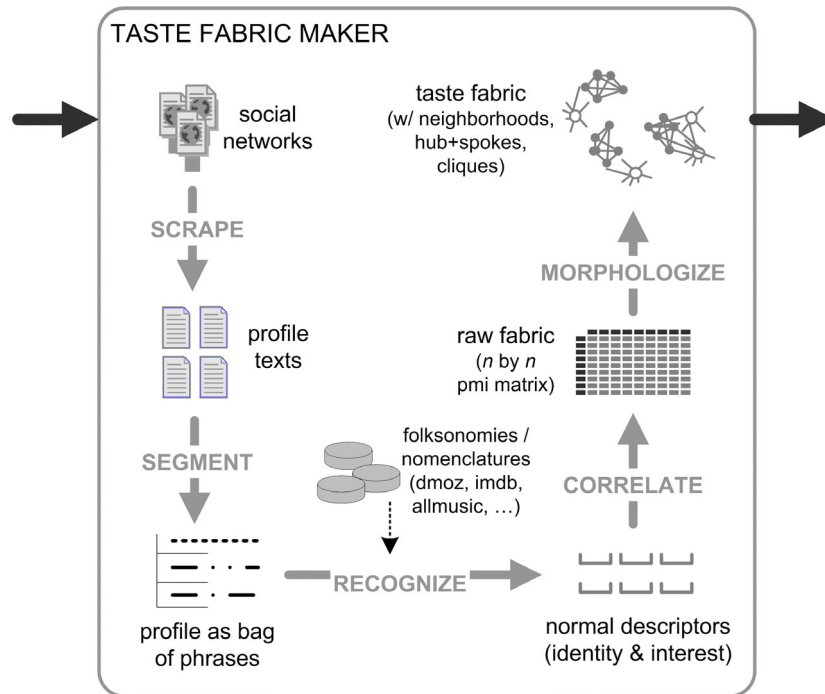


Figure 2: Implemented architecture of the taste fabric maker

3.1 Acquiring profiles from social networking sites

The present implementation of the taste fabric sources from a one-time crawl of two web-based social network sites, which took place over the course of six months in 2004. The shortcoming of this approach is that we were only able to mine 100,000 personal profiles from the sites, and only approximately 80% of these profiles contained substantive content because about 20% of users elected to not make their profile details publicly visible to our robotic crawler. Also, one-time crawls prevent us from being able to engage in more interesting dynamic tracking of profiles which would potentially allow us to animate taste fabrics through time. At press time, we are in discussions with two social network websites to gain research access to their user profiles, which should allow for the implementation that we discuss in this paper to be deployed on a larger scale.

At every step of mining, we are careful not to compromise the anonymity or personal information of social network users. In fact, in the end product, all traces of individual users, as well as their idiosyncratic speech, is cleaned out of the taste fabric. From our 100,000 seed profiles, we maintained only the text of the categorical descriptors (e.g. “music”, “books,” “passions/general interests”) and none of the personal information including names and screen names. We chose

two social networks rather than one, to attempt to compensate for the demographic and usability biases of each. One social network has its membership primarily in the United States, while the other has a fairly international membership. Both however, have nearly identical descriptor categories, and both sites elicit users to specify punctuation-delimited descriptors rather than sentence-based descriptors. One cost to mining multiple social networks is that there is bound to be some overlap in their memberships (by our estimates, this is about 15%), so these twice-profiled members may have disproportionately greater influence on the produced fabric.

3.2 Segmenting profiles

Once profile texts are acquired, these texts need to be segmented. First, texts are easily segmented based on their interest categories. Recall in Figure 1 that texts are distributed across templated categories, e.g., passions/general interests, books, music, television shows, movies, sports, foods, “about me.” Experience with the target social network websites tell us that most users type free-form natural language text into “about me,” and natural language fragments for the specific interest categories. For the passions / general interest category, text is likely to be less structured than for specific interest categories, but still more structured than “about me.” Perhaps this is due to the following psychology—for specific interests, it is clear what the instances would be, e.g. film names, director names, and film genres for the films category, yet for the general interest category, the instance types are more ambiguous—so that field tends to elicit more idiosyncratic speech.

For each profile and category, its particular style of delimitation is heuristically recognized, and then applied. Common delimitation strategies were: comma-separated, semicolon-separated, stylized character sequence-separated (e.g. “item 1 \./ item 2 \./ ...”), new line –separated, commas with trailing ‘and’, and so on. Considering a successful delimitation as a category broken down into three or more segments, approximately 90% of specific categories were successfully delimited, versus about 75% of general categories. We did not attempt to segment “about me.” Unsegmentable categories were discarded.

3.3 Ontology-driven natural language normalization

After segmentation, descriptors are normalized by mapping them into a formal ontology of identity and interest descriptors (Figure 3). Newly segmented profiles are represented as lists containing casually-stated natural language fragments referring to a variety of things. They refer variously to authorships like a book author, a musical artist, or a filmmaker; to genres like “romance novels,” “hip-hop,” “comedies,” “French cuisine”; to titles like a book’s name, an album or song, a television show, the name of a sport, a type of food; or to any combination thereof, e.g. “Lynch’s Twin Peaks,” or “Romance like Danielle Steele.” To further complicate matters, sometimes only part of an author’s name or a title is given, e.g. “Bach,” “James,” “Miles,” “LOTR,” “The Matrix trilogy.” Then of course, the items appearing under the general interest categories can be quite literally anything.








TASTE FABRIC'S INSTANCE TYPES		
category	types	ontology sources
 identities	subculture, __ lover, taste echelon	wikipedia's "list of subcultures", dmoz
 films	filmmaker, film title, film genre	imdb, dmoz
 books	author, title, genre	amazon, wikipedia, dmoz
 music	artist, album, song, genre/decade	allmusic, amazon, dmoz
 foods	dish name, ethnicity, ingredient, course	allrecipes, foodsubs
 sports	name, genre	dmoz, amazon
 television	show name, genre	tvguide's "showguide", dmoz

Figure 3: Table of instance type ontology and data sources

Figure 3 presents the ontology of descriptor instance types for the present taste fabric. At the top-level of the ontology are six specific interest categories plus one general interest category (i.e., "identities"). Also, as shown, there are roughly 25 second-level ontological types. There are a total of 21,000 recognizable interest descriptors, and 1,000 recognizable identity descriptors, sourcing from ontologies either scraped or XML-inputted from The Open Directory Project (dmoz)¹, the Internet Movie Database (imdb)², TV Tome³, TV Guide⁴, Wikipedia⁵, All Music Guide⁶, AllRecipes⁷, and The Cook's Thesaurus⁸. Figure 3 only lists the primary sources, and lists them in order of descending saliency. The diversity and specificity of types ensures the maximal recognition capability over the free-form natural language in the profiles.

Ontologizing identity. The ontology of 1,000 identity descriptors required the most intensive effort to assemble together, as we wanted them to reflect the types of general interests talked about in our corpus of profiles; this ontology was hand-engineered out of a few nomenclatures and folksonomies—most prominently Wikipedia's extensive list of subcultures and The Open Directory Project's hierarchy of subcultures and hobbies. We also generated identity descriptors in the form "(blank) lovers" where blank was replaced with major genres in the rest of our ontology, e.g. "book lovers," "country music lovers," etc. Some profiles simply repeat a select subset of interest descriptors in the identity descriptors category, so having the "(blank) lovers"

¹ <http://www.dmoz.org>

² <http://www.imdb.com>

³ <http://tv Tome.com>

⁴ <http://tvguide.com>

⁵ <http://www.wikipedia.org>

⁶ <http://www.allmusic.com>

⁷ <http://allrecipes.com>

⁸ <http://www.foodsubs.com>

template would facilitate the system in recognizing these examples. The mapping from the general interest category into the identity descriptors ontology is far more indirect a task than recognizing specific interests because the general interest category does not insinuate a particular ontology in its phrasing. Thus, to facilitate indirect mapping, each identity descriptor is annotated with a bag of keywords which were also mined out from Wikipedia and The Open Directory Project—so for example, the “Book Lover” identity descriptor is associated with, *inter alia*, “books,” “reading,” “novels,” and “literature.” Because we employed two parallel mechanisms for identity descriptors, i.e. cultures versus “(blank) lovers,” we cannot be completely assured that these do not overlap—in fact, they are known to overlap in a few cases, such as “Book Lovers” and “Intellectuals” or “Indie Rock Music Lovers” (genre of music) and “Indie” (subculture). Most cases of overlap, however, are much more justified because the cultural lexicon, just as natural language, cannot be flattened to a canon. Perhaps the most controversial interpretative choice we made was—for the sake of bolstering recognition rates—up-casting descriptors until they could be recognized in the identity ontology. For example, while “Rolling Stones” is not in the ontology of identity descriptors, we automatically generalize it until it is recognized, or all generalizations are exhausted—in the case of “Rolling Stones,” it is up-cast into “Classic Rock Music Lovers.”

Popularity-driven disambiguation. To assist in the normalization of interest descriptors, we gathered aliases for each interest descriptor, and statistics on the popularity of certain items (most readily available in The Open Directory Project) that the system uses for disambiguation. For example, if the natural language fragment says simply “Bach,” the system can prefer the more popular interpretation of “JS Bach” over “CPE Bach.”

Situated semantic recognition. Once a profile has been normalized into the vocabulary of descriptors, they are relaxed semantically using a spreading activation (Collins & Loftus, 1975) strategy over the formal ontology, because more than simply being flat wordlists, the ontological instances are cross-annotated with each other to constitute a fabric of metadata. For example, a musical genre is associated with its list of artists, which in turn is associated with lists of albums, then of songs. A book implies its author, and a band implies its musical genre. Descriptors generated through metadata-association are included in the profile, but at a spreading discount of 0.5 (read: they only count half as much). This ensures that when an instance is recognized from free-form natural language, the recognition is situated in a larger semantic context, thus increasing the chances that the correlation learning algorithm will discover latent semantic connections.

In addition to popularity-driven disambiguation of, e.g. “Bach” into “JS Bach,” we also several other disambiguation strategies. Levenshtein (1965/1966) edit distance is used to handle close misspellings such as letter deletions, consecutive key inversions, and qwerty keyboard near-miss dislocations, e.g. “Bahc” into “Bach.” Semantically empty words such as articles are allowed to be inserted or deleted for fuzzy matching, e.g. “Cardigans” into “The Cardigans” (band).

Using this crafted ontology of 21,000 interest descriptors and 1,000 identity descriptors, the heuristic normalization process successfully recognized 68% of all tokens across the 100,000 personal profiles, committing 8% false positives across a random checked sample of 1,000 mappings. Here, “tokens” refers to the natural language fragments outputted by the segmentation process; a recognition is judged successful if after stripping away semantically empty words, the

token finds correspondence with an instance in the ontology, while remaining within the heuristically-specified tolerances for misspelling and popularity-driven disambiguation. We suggest that this is a good result considering the difficulties of working with free text input, and enormous space of potential interests and identities.

3.4 Correlation: weaving the raw fabric

From the normalized profiles now each constituted by normalized identity and interest descriptors, correlation analysis using classic machine learning techniques reveals the latent semantic fabric of interests, which, operationally, means that the system should learn the overall numeric strength of the semantic relatedness of every pair of descriptors, across all profiles. In the recommender systems literature, our choice to focus on the similarities between descriptors rather than user profiles reflects an item-based recommendation approach such as that taken by Sarwar *et al.* (2001).

Technique-wise, the idea of analyzing a corpus of profiles to discover a stable network topology for the interrelatedness of interests is similar to how *latent semantic analysis* (Landauer, Foltz & Laham, 1998) is used to discover the interrelationships between words in the document classification problem. For our task domain though, we chose to apply an information-theoretic machine learning technique called *pointwise mutual information* (Church & Hanks, 1990), or PMI, over the corpus of normalized profiles. For any two descriptors f_1 and f_2 , their PMI is given in equation (1). The probability of a descriptor, $\Pr(f_i)$, is defined here as the frequency of global occurrences of f_i divided by the summed frequency of global occurrences for all descriptors.

$$PMI(f_1, f_2) = \log_2 \left(\frac{\Pr(f_1 f_2)}{\Pr(f_1) \Pr(f_2)} \right) \quad (1)$$

Looking at each normalized profile, the learning program judges each possible pair of descriptors in the profile as having a correlation, and updates that pair's PMI. What results is a 22,000 x 22,000 matrix of PMIs, because there are 21,000 interest descriptors and 1,000 identity descriptors in the ontology. After filtering out descriptors which have a completely zeroed column of PMIs, and applying thresholds for minimum connection strength, we arrive at a 12,000 x 12,000 matrix (of the 12,000 descriptors, 600 are identity descriptors), and this is the raw interest fabric. This is too dense to be visualized as a semantic network, but we have built less dense semantic networks by applying higher thresholds for minimum connection strength, and this is the reason why clustering seem to appear in the InterestMap taste fabric visualization.

Criticism and limitations. A common critique heard about our approach is one that questions the efficacy of using the PMI metric for association. It has been suggested that we should look at collocations of greater rank than binary. Following our initial InterestMap publication, we extended the work by using morphological opening plus thresholding, as is done in image processing, to try to discover larger blocks of collocations which we call *neighborhoods*. This is to be discussed imminently. Additionally, another suggestion we are considering at press is negative collocation, that is, the collocation of a descriptor's absence with other descriptors. This would address an apparent flaw of pointwise mutual information, which is that it "overvalues frequent forms" (Deane, 2005), and would shed a new interpretation on the Semiotician

Ferdinand de Saussure’s structuralist enunciation that meaning must be ‘negatively defined’ (1915/1959).

3.5 Looking at topological features

The raw fabric has two extant topological features worthy of characterization—*identity hubs* and *taste cliques*. In addition, we describe what we believe to be a novel application of mathematical morphology (Serra, 1982; Haralick, Sternberg & Zhuang, 1987) in conjunction with spreading activation (Collins & Loftus, 1975) to discover the taste neighborhoods we hinted at in Section 1.

Identity hubs behave like seams in the fabric. Far from being uniform, the raw fabric is lumpy. One reason is that identity hubs “pinch” the network. Identity hubs are *identity descriptor nodes* which behave as “hubs” in the network, being more strongly related to more nodes than the typical *interest descriptor node*. They exist because the ontology of identity descriptors is smaller and less sparse than the ontology of interest descriptors; each identity descriptor occurs in the corpus on the average of 18 times more frequently than the typical interest descriptor. Because of this ratio, identity hubs serve an *indexical* function. They give organization to the forest of interests, allow interests to cluster around identities. The existence of identity hubs allows us to generalize the granular location of what we are in the fabric, to *where in general we are* and what identity hubs we are closest to. For example, it can be asked, what kinds of interests do “Dog Lovers” have? This type of information is represented explicitly by identity hubs.

Taste cliques as agents of cohesion. More than lumpy, the raw fabric is denser in some places than in others. This is due to the presence of *taste cliques*. Visible in Figure 5, for example, we can see that “Sonny Rollins,” is straddling two cliques with strong internal cohesion. While the identity descriptors are easy to articulate and can be expected to be given in the special interests category of the profile, tastes are often a fuzzy matter of aesthetics and may be harder to articulate using words. For example, a person in a Western European taste-echelon may fancy the band “Stereolab” and the philosopher “Jacques Derrida,” yet there may be no convenient keyword articulation to express this. However, when the taste fabric is woven, cliques of interests seemingly governed by nothing other than taste clearly emerge on the network. One clique for example, seems to demonstrate a Latin aesthetic: “Manu Chao,” “Jorge Luis Borges,” “Tapas,” “Soccer,” “Bebel Gilberto,” “Samba Music.” Because the cohesion of a clique is strong, *taste cliques* tend to behave much like a singular identity hub, in its impact on network flow. In the following Section, we discuss how InterestMap may be used for recommendations, and examine the impact that identity hubs and taste cliques have on the recommendation process.

3.6 Carving out taste neighborhoods with mathematical morphology

From the raw fabric, another step of processing is needed to reveal *taste neighborhoods*—patches of taste cohesion that are larger than taste cliques and more stable than ephemeral taste ethoi. Taste neighborhoods of course, overlap with one another, and the discovery and definition of taste neighborhoods seems even prone to the Ptolemaic dilemma—some nodes must be designated as “center of the universe,” and the choice of these centric nodes can greatly affect the resultant neighborhood definition. Two taste neighborhood with different Ptolemaic centers are shown in Figure 4.

Taste ethos from spreading activation. While the technical specifics for the discovery process are potentially lengthy, we sketch a conceptual overview of the implementation here. The raw n by n correlation matrix is re-viewed as a classic spreading activation network (Collins & Loftus, 1975). That is to say, activation spreads outward from an origin node to all the connected nodes, then from all connected nodes to each of their connected nodes. The obvious observation here is that in our correlation situation, all nodes are connected to a large percentage of the graph, so our graph is super connected. However, what makes the spreading activation meaningful is that the strength of the spread activation is proportional to the strength of the PMI along any edge in the graph. The energy of the spreading is also inhibited as the number of hops away from the origin grows, according to a per hop discount rate (e.g. 50%) So, spreading with a low tolerance (or, a high threshold for activation), and outward from “Jazz,” “Yoga” (two-hops away) are reachable, but the energy attenuates before the “Football” (also, two-hops away) node can be activated.

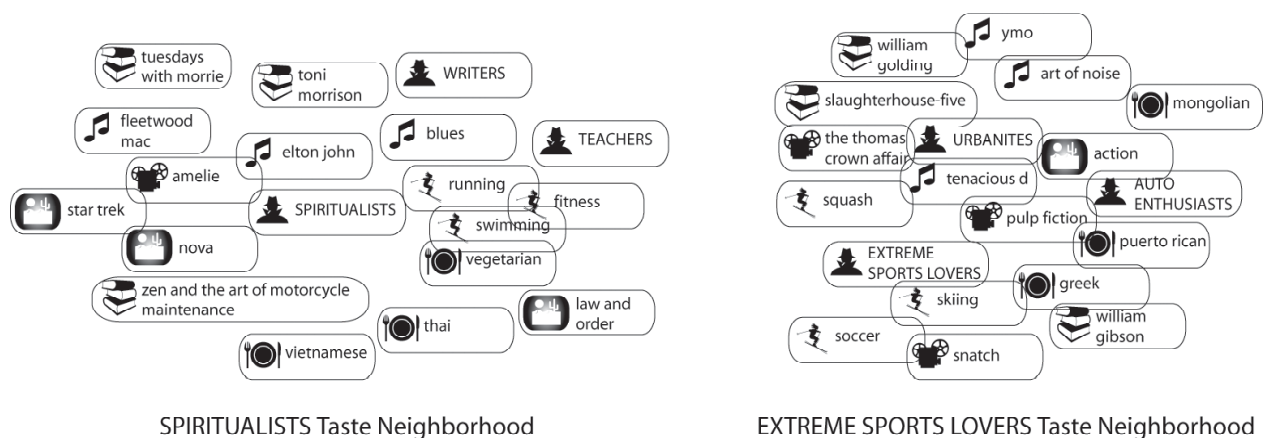


Figure 4: Two Ptolemaically-centered taste neighborhoods, computer generated with the follow parameters—a maximum of 50 nodes in each neighborhood, up to the first three instances of any category type are shown. Spatial layout is not implied by the neighborhood, so the above are manually arranged.

Spreading activation outward from an origin node, the result can be likened to that node’s defeasible (default, in the absence of other inputs or biases) taste ethos. This taste ethos is too small when spreading activation is configured with a modest tolerance. On the other hand, if the tolerance is increased too dramatically, the taste ethos will grow in size but its stability will be undermined due to a well-known problem in graph-based inference—beginning at two hops away, reached nodes lose their semantic relevance to the origin node very rapidly. Think of this as the *telephone game effect*—playing the childhood game of telephone, the first couple of hops are still recognizable, but recognition often rapidly tapers off after the first couple of hops. The effect is also observed by Google in their PageRank algorithm (Brin & Page, 1998) for scoring the salience of web pages by taking a voting scheme of salience. They noted that high-rank pages tended to link to high-quality pages, and those to other high-quality pages, but after distance=2, reliability tapered off rapidly.

Mathematical morphology. To discover neighborhoods of taste which are larger than particular node-centric ethoi, but which are still nonetheless stable, we borrow two techniques from the field of mathematical morphology (Serra, 1982; Haralick, Sternberg & Zhuang, 1987) and that are widely used in the image processing literature which appropriates them for object

segmentation – morphological opening and thresholding. Morphological opening is the mathematical composition of two operators *erosion* and *dilation*, in that order. The intuition is that erosion ‘eats away’ at the boundaries of an object, whereas dilation ‘grows’ the boundaries of the object. However, erosion and dilation are not inverses because both erosion and dilation are injective, that is, they are many-to-one and lossful transformations. The effect of morphological opening is also quite intuitive—it removes small objects ‘disturbances’ and opens up gaps when they are located near a boundary. There is morphological opening and there is also morphological closing which is dilation composed with erosion; closing fills in holes and around boundaries more than opening. We employ opening because it is a bit crisper. Opening eliminates blemishes while closing magnifies blemishes. The other technique, thresholding, is frequently used to post-process an opened image. Applying a fixed threshold to an opened image simply turns every pixel above the threshold to 1, and below the threshold to 0.

Erosion and dilation over spread activations. We choose identity nodes as the centric origins for spreading because they are in general more stable places to start from. This follows the rationale that identities are stronger cultural fixtures than a book or a music album, generally speaking. From the identity nodes, we apply a relatively lenient discount, e.g. 0.8, and spread to define a fairly relevant neighborhood. This is repeated over all identity nodes, begetting an ethos for each identity node. Where ethoi overlap, the max of the node’s energy is taken, rather than the sum of the node’s energies. Now, an erosion is applied, trimming back the weakest boundary nodes, followed by a dilation, growing the boundary by adding some energy to all nodes connected to the boundary, pushing some of them over the activation threshold and thus growing the mass. In the current implementation, two iterations of opening are performed, though the meaning of this varies widely with the choice of thresholds and other considerations.

In this manner, larger stable masses of nodes, termed taste neighborhoods, are discovered. Thresholding can help us trim a neighborhood to an arbitrary node-size. For visualizations such as InterestMap, neighborhoods comprised of up to thirty nodes seem visually appropriate. We believe that the application of morphological opening and thresholding to a spreading activation network in order to discover larger stable neighborhoods is a novel use, though we do not evaluate this claim within this paper’s scope.

Summary. This section discussed an implementation of weaving the interest fabric out of social networks. Profiles mined from two social network websites were heuristically segmented and normalized according to a heterogeneous ontology assembled together from a variety of data sources. After normalization, correlation analysis learned the affinities between descriptors, and mathematical morphology over the “raw” fabric enabled taste neighborhoods to be discovered and overlaid onto the fabric. Next, we demonstrate the semantic uses of the taste fabric within application contexts.

4 What Is a Taste Fabric Good for?

As a rich tapestry of interconnected interests and identities, the taste fabric brings us closer to improvisational manipulations of meaning, and affords us at least three semantic functions— the creation of semantically flexible user representations, cross-domain taste-based recommendation, and the computation of taste-similarity between people. This section explores these three basic

semantic functions in the context of a survey of three applications we have developed. InterestMap is a taste-based recommendation system that leverages interactive visualization of neighborhoods to make the recommendation mechanism transparent, thereby enhancing users' trust perceptions of the recommender. Ambient Semantics uses the taste fabric to facilitate social introductions between two strangers, based on their shared taste. IdentityMirror is a digital mirror for identity self-management. Whereas a real mirror shows you what you look like, IdentityMirror shows you who you are. It explores semantically flexible user representations by allowing time, orderliness, and current events in the world to nuance the representation of the viewer.

4.1 InterestMap

InterestMap (Liu & Maes, 2005a) visualizes the topology of the taste fabric, and in particular it depicts taste cliques, identity hubs, and taste neighborhoods as a navigatable map. As shown in Figure 5a, users can browse InterestMap's tapestry of neighborhoods, cliques and identity hubs, or, as depicted in Figure 5b, they can interactively build up their own taste ethoi, by searching for and attaching descriptors to a stationary "who am i?" node. The act of connecting a descriptor to the self is deeper than making a mere superficial keyword association since each descriptor is actually something more like a semantic cloud. Once a user has connected several descriptors to his self, those semantic clouds begin to intersect, overlap, and mingle. They begin to imply that other descriptors, which the user has not selected himself, *should* be within the user's taste. Hence the notion of a visual recommendation.

Taste-based recommendation. InterestMap can, given a profile of the user's interests, recommend in a cross-domain way, books, songs, cuisines, films, television shows, and sports to that user based on taste. The user's interests are normalized according to aforementioned processes and mapped into the taste fabric. These nodes in the fabric constitute a particular activation configuration that is unique to the user, and the total situation described by this configuration is the fuzzy taste model of the user. To make recommendations, activation is spread outward from this configuration, into the surrounding nodes. Some nodes in the surrounding context will be activated with greater energy because they are more proximal to the taste signature of the starting configuration. The nodes activated with the highest energy constitute the user's recommendation. Figure 5b shows a visualization of the recommendation process. The user's self-described interests are the descriptors directly connected to the "who am i?" node. Each of these interests automatically entail other strongly connected descriptors. This is visually expressed well in the InterestMap visualization because a node getting pulled toward "who am i?" will tug a whole web of nodes behind it. Since the visualization starts with just the "who am i?" node visible on the screen, specifying just a couple of interests can literally fill up the screen with its taste entailments. To visualize the spreading activation mechanism, the size and yellowness of nodes diminishes as activation spreads outward from the "who am i?" node.



Figure 5: Two screenshots of the InterestMap interactive visualization. **5a** (top) depicts a user browsing neighborhoods of taste visually. **5b** (bottom) depicts a user visualizing his own taste ethos by dragging and connecting interesting nodes to the “who am i?” node.

Visual recommendation enhances transparency and trust. That a user trusts the recommendations served to him by a recommender system is important if the recommender is to be useful and adopted. Among the different facilitators of trust, Wheelless & Grotz (1977) identify transparency as a prominent desirable property. When a human or system agent discloses its assumptions and reasoning process, the recipient of the recommendation is likely to feel less apprehensive toward the agent and recommendation. Also in the spirit of transparency, Herlocker, Konstan & Riedl (2000) report experimental evidence to suggest that recommenders

which provide explanations of its workings experience a great user acceptance rate than otherwise.

Unlike opaque statistical mechanisms like collaborative filtering (Shardanand & Maes, 1995), InterestMap's mechanism for recommendation can be communicated visually. The idiosyncratic topology of this taste fabric symbolizes the common taste tendencies of a large group of people. For example, in Figure 5a, it is plain to see that "Sonny Rollins" and "Brian Eno" are each straddling two different cliques of different musical genres. The rationale for each recommendation, visually represented as the spreading of flow across the network, is easily intelligible. Thus it may be easier for a user to visually contextualize the reasons for an erroneous recommendation, *e.g.* "I guess my off-handed taste for Metallica situated me in a group of metal heads who like all this other stuff I hate."

Although we have not yet implemented such a capability, the ability to interact with the InterestMap network space would also afford the system an opportunity to *learn* more intelligently from user feedback about erroneous recommendations. Rather than a user simply stating that she did not like a particular recommendation, she could black out or deprecate particular clusters of the network which she has diagnosed as the cause of the bad recommendation, *e.g.* "I'll black out all these taste cliques of heavy metal and this identity hub of "Metal Heads" so the system will not make *that* mistake again."

4.2 Ambient Semantics

Ambient Semantics (Maes *et al.*, 2005) is a wearable contextual information system that supports users in discovering objects and meeting people through pithy *just-in-time feedback* given in the crucial first moments of an encounter. Here is an example of a use case involving the discovery of a new book: Wearing the Ambient Semantics RFID reader wristband, you pick up a copy of Marvin Minsky's "Society of Mind" book. Through your cell phone display, the system tells you that you would be particularly interested in section 3 because it is relevant to your current research topics. It would tell you that your friends Henry and Barbara listed this book among their favorites, and that the author's expressed opinions seem sympathetic to your own, based on semantic analyses of both your writings. The system can indicate that you would find the book tasteful because it can use taste fabric to detect that it is indeed within close proximity to your taste ethos, translating to a strong taste-based recommendation.

Exposing shared taste-context between two strangers. The second use case concerns the system facilitating social introductions by breaking the ice. This scenario demonstrates using the taste fabric for the quantification and qualification of the taste-similarity between two strangers. First, a scenario. You are at a business networking event where Ambient Semantics wristwatches have been given to the attendees. You are tired of the same old conversation starters – what's your name – who do you work for – how do you like it here? – so you head to the Ambient Semantics kiosk where people are meeting each other in a new way. You introduce yourself to a lady standing next to you. By virtue of your handshake, the physical surroundings are transformed. The music and lighting in the area changes to suit the shared aspects of yours and the lady's tastes. Some graphics of kayaking are thrown up on the kiosk display, as well as the faces of some people. The lady says to you, 'so you know Bob and Terry too? Are you in the Boston Outdoor Society too?'

Calculating taste-similarity: quantitatively vs. qualitatively. There is more than one good way to use taste fabric to calculate the taste-similarity of two people. The more direct way is to measure the intersection of two spread activations. Taking each person's seed profile of interests and mapping it into the taste fabric, we arrive at an initial configuration. Spreading activation outward from this configuration defines a semantic neighborhood, which earlier in the paper we referred to as a person's taste ethos. Taking the semantic intersection of two or more persons' ethoi, we arrive at the quantitative calculation of taste-similarity.

However, another intriguing possibility is to make a qualitative calculation about taste-similarity. Although the intersection of two taste ethoi is mathematically satisfying, it is not easily explainable and articulated. In other words, having the system explain that "the two of you share taste because you both have interests x, y, and z in your spreading activation clouds" is inappropriate. More articulate would be to cite a shared habitation of taste neighborhoods, for example, this explanation—"the two of you share taste because both of you are adventurers and lovers of wine." Here, the mechanism of the recommendation feels more transparent. To calculate qualitative similarity, each person's taste ethos would be used to score the degree of a person's habitation across the various taste neighborhoods, which as you recall, are centered around identity nodes. Like the classic k-nearest neighbors classification scheme, here we classify persons by their k-nearest taste neighborhoods. Having completed this mapping, the subset of neighborhoods shared among the two or more persons become those persons' shared situation. To communicate shared neighborhoods to the persons, the neighborhoods could be effectively visualized on a screen, or, neighborhoods are safely summarized by stating the identity nodes which live within that neighborhood.

4.3 IdentityMirror

What if you could look in the mirror and see not just what you look like, but also who you are? Identity mirror (Figure 6) is an augmented evocative object that reifies its metaphors in the workings of an ordinary mirror. When the viewer is distant from the object, a question mark is the only keyword painted over his face. As he approaches to a medium distance, larger font sized identity keywords such as "fitness buffs", "fashionistas", and "book lovers" identify him. Approaching further, his favorite book, film, and music genres are seen. Closer yet, his favorite authors, musicians, and filmmakers are known, and finally, standing up close, the songs, movies, and book titles become visible.

The Identity Mirror learns and visualizes a dynamic model of a user's identity and tastes. Looking into it, the viewer's face is painted over with identity and keywords, sourced from this dynamic user model. Taste fabric is used to interpret an initial seed profile into a semantic situation within the fabric. For instance, the viewer specifies that he listens to "Kings of Convenience" and enjoys the fiction of Vladimir Nabakov, and using this, taste fabric situates the viewer within its multiple neighborhoods of taste. The keywords which paint over the viewer's face represent his context within taste-space.

5 Advanced Discussion

In this section, we present an evaluation of the taste fabric, present related work, and discuss other ways in which this work is of consequence to the semantic mining and Semantic Web communities.

5.1 Evaluation

We evaluate the quality of the taste fabric apropos a *telos* of recommendation, scrutinizing the performance of recommending interests via spreading activation over the taste fabric, as compared with a classic collaborative filtering recommender. Much of this discussion is adapted from (Liu & Maes, 2005a).

In this evaluation, we introduced three controls to assess two particular features: 1) the impact that identity hubs and taste cliques have on the quality of recommendations; and 2) the effect of using spreading activation rather than a simple tally of PMI scores. Notably absent is any evaluation for the quality of the produced taste neighborhoods, because here we consider only quantitative and not qualitative recommendation. Qualitative recommendation is not claimed to outperform quantitative recommendation in terms of accuracy—our suggestion was that linguistically identifying and visually illustrating two persons’ cohabitations of taste neighborhoods should facilitate trust and transparency in the recommender’s process.

In the first control, identity descriptor nodes are simply removed from the network, and spreading activation proceeds as usual. In the second control, identity descriptor nodes are removed, and n -cliques⁹ where $n > 3$ are weakened¹⁰. The third control does not do any spreading activation, but rather, computes a simple tally of the PMI scores generated by each seed profile descriptor for each of the 11,000 or so interest descriptors. We believe that this successfully emulates the mechanism of a typical non-spreading activation item-item recommender because it works as a pure information-theoretic measure.

We performed five-fold cross validation to determine the accuracy of the taste fabric in recommending interests, versus each of the three control systems. The corpus of 100,000 normalized and metadata-expanded profiles was randomly divided into five segments. One-by-one, each segment was held out as a test corpus and the other four used to train a taste fabric using PMI correlation analysis. The final morphological step of neighborhood discovery is omitted here.

Within each normalized profile in the test corpus, a random half of the descriptors were used as the “situation set” and the remaining half as the “target set.” Each of the four test systems uses the situation set to compute a *complete recommendation*— a rank-ordered list of all interest descriptors; to test the success of this recommendation, we calculate, for each interest descriptor in the target set, its percentile ranking within the complete recommendation list. As shown in (2),

⁹ a qualifying clique edge is defined here as an edge whose strength is in the 80th percentile, or greater, of all edges

¹⁰ by discounting a random 50% subset of the clique’s edges by a Gaussian factor (0.5μ , 0.2σ).

the overall accuracy of a complete recommendation, $a(CR)$, is the arithmetic mean of the percentile ranks generated for each of the k interest descriptors of the target set, t_i .

$$a(CR) = \frac{1}{k} \sum_{i=1}^k \text{percentile}(t_i, CR) \quad (2)$$

We opted to score the accuracy of a recommendation on a sliding scale, rather than requiring that descriptors of the target set be guessed exactly within n tries because the size of the target set is so small with respect to the space of possible guesses that accuracies will be too low and standard errors too high for a good performance assessment. For the TASTEFABRIC test system and control test systems #1 (Identity OFF) and #2 (Identity OFF and Taste WEAKENED), the spreading activation discount was set to 0.75). The results of five-fold cross validation are reported in Figure 7.

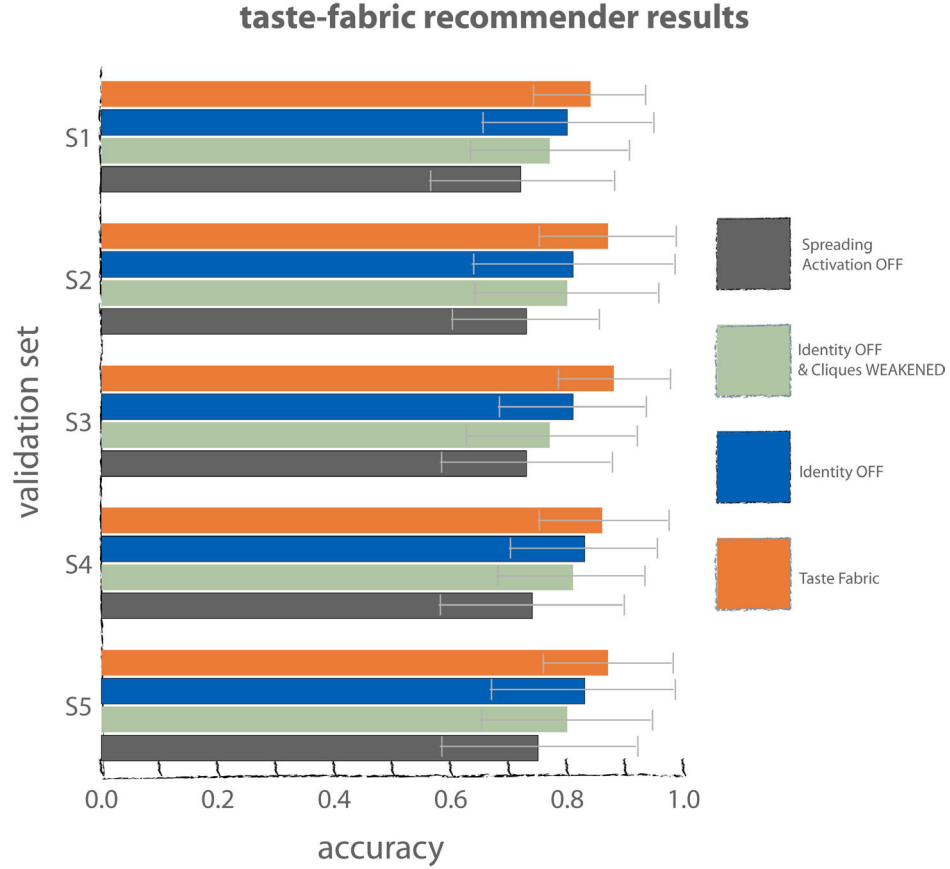


Figure 7. Results of five-fold cross-validation of taste-fabric recommender and three control systems on a graded interest recommendation task.

The results demonstrate that on average, the full taste fabric recommended with an accuracy of 0.86. In control #1, removing identity descriptors from the network not only reduced the accuracy to 0.81, but also increased the standard error by 38%. In control #2, removing identity descriptors and weakening cliques further deteriorated accuracy slightly, though insignificantly, to 0.79. When spreading activation was turned off, neither identity hubs nor taste cliques could

have had any effect, and we believe that is reflected in the lower accuracy of 73%. However, we point out that since control #3's standard error has not worsened, its lower accuracy should be due to overall weaker performance across all cases rather than being brought down by exceptionally weak performance in a small number of cases.

We suggest that the results demonstrate the advantage of spreading activation over simple one-step PMI tallies, and the improvements to recommendation yielded by identity and taste influences. Because activation flows more easily and frequently through identity hubs and taste cliques than through the typical interest descriptor node, the organizational properties of identity and taste yield proportionally greater influence on the recommendation process; this of course, is only possible when spreading activation is employed.

5.2 Related works

A cultural metadata approach to musical taste. Whitman and Lawrence (2002) developed a metadata model for characterizing the taste coherence of musical genres. Mining adjectival and noun phrases collocated with musical artist discussions in newsgroups and chatrooms, they applied machine learning to automatically annotate music artists with what they termed “community metadata.” Then Whitman and Smaragdis (2002) applied community metadata to build cultural signatures for music genres that could be used, in conjunction with the auditory signal, to classify unknown artists based on style similarity. Their notion of a metadata signature for musical styles is sympathetic to our notion of taste ethos and taste neighborhood, and both systems take a bottom-up metadata-driven view of meaning definition. A chief difference between our two works is that taste knowledge is located in descriptive word-choice in their system (e.g. “wicked,” “loud”), and located in interest-choice in our system, that is, the choices of what people consume (e.g. “Britney Spears”, “Oreo cookies”).

Social information filtering. In prior work, one of the authors co-developed a well-known technique for item recommendation based upon nearest taste-neighbor, the approach known variously as social filtering, or collaborative filtering. Shardanand and Maes (1995) represent users as vectors of (item, rating) pairs, and compute taste-similarity as statistical correlation between the two vectors, or alternatively as cosine similarity of the two vectors in n -dimensional item space. In their Ringo social music browsing system, users are recommended a list of potential ‘tastemates’ on the basis of taste-similarity. One difference between our two approaches is that social filtering maintains distinct user profiles, whereas taste fabrics dissolves user boundaries, and is, in their terminology, a ‘content-based filtering’ approach. In distilling a reusable knowledge resource out of social network profiles that can be reappropriated for a variety of other purposes not concerned with the original social network community, it is necessary to protect the privacy of the original users, and we suggest that taste fabrics serves as a model for doing so. Also relevant is Sarwar *et al.*'s (2001) item-based collaborative filtering approach to recommendation, which, like taste fabrics, relies upon item-item correlation rather than user-user correlation. Taste fabric exceeds item-based filtering by use of extensive metadata to ‘relax’ the meaning from the item itself, by defining identity descriptors as supernodes, and by representing users as k -nearest neighborhoods. In general, collaborative filtering is more representational opaque whereas spreading activation over neighborhoods can be visualized and more easily debugged.

Social network analysis and relational mining. Much research has examined the explicit structure of social networks, and studied their topologies via graph theory. Newman (2001) mined scientific coauthorship networks and found that collaborations ‘funneled’ through gatekeeper scientists. In taste fabrics, identity hubs, and hubs created around particularly salient interest descriptors constitute a similar topological feature. Jensen and Neville (2002) mined structured metadata relations from the Internet Movie Database (imdb.com) called ‘schema’ and learned a Bayesian network model to represent and predict item distances probabilistically. They also model the relational semantics of social network relations implied between movie actors from the Internet Movie Database and the Hollywood Stock Exchange (www.hsx.com). Finin *et al.* (2005) examine how the FOAF (“friend-of-a-friend”) ontology applies Semantic Web concepts to enable efficient exchange of and search over social information, illustrating how social networks could develop with its semantics already explicit. Finally one work which considers the semantic content entailments of social network users is McCallum, Corrada-Emmanuel, and Wang’s (2005) modeling of Author-Recipient-Topic correlations in a social network messaging system. Given the topic distributions of email conversations, the ART model could predict the role-relationships of author and recipient. The work consider group clusters and dyadic relationship dynamics but does not consider cultural aggregates as is the concern of our present work.

Large-scale commonsense knowledge networks. Taste fabrics are a rich tapestry which define the meaning space of taste and interests. They are represented as semantic networks and reasoning is performed via spreading activation over this network. This approach to knowledge representation and reasoning builds upon previous work in large-scale semantic knowledge bases such as WordNet (Fellbaum, 1998) and ConceptNet (Liu & Singh, 2004). WordNet is a semantic network whose nodes are words, and edges are various nymic lexical relations between the words, e.g. a “dog” has the hypernym of “canine.” ConceptNet is a semantic network of commonsense knowledge whose 200,000 nodes are verb phrases (“eat burger”, “take shower”), and 1.6 million edges are one of 20 kinds of world semantic relations (e.g. “EffectOf,” “PropertyOf,” “DesireOf”), e.g. (EffectOf “be hungry” “cook food”). ConceptNet and taste fabrics reason similarly by activating a seed configuration of nodes, and spreading activation outward to define a semantic context. Both resources are densely connected, semantically extensive within their respective domains, and allow for improvisational manipulations of meaning to take place atop it.

5.3 Reusable Methodologies

Sanitary semantic mining. The *sanitariness* of a mined knowledge resource is the degree to which it is purged of idiosyncrasy, especially idiosyncratic traces of user-specific information, and also idiosyncrasies which implicate the original application domain from which the resource was mined. When a knowledge resource is sanitary, assurances can be made that private user data is not recoverable, and that the resource is sufficiently context-free so that it could potentially be used to solve problems across a variety of domains. Taste fabrics are an illustration of how a sanitary knowledge resource can be mined out of a highly idiosyncratic and application specific data source such as self-descriptive social network profiles. Because it is sanitized, taste fabrics can be publicly distributed and used to power applications living in other domains.

When mining social network data, concern for privacy and copyrights of user data make derivative works especially problematic; yet there is a great need and opportunity to infer valuable semantic knowledge from these sources. Ensuring data anonymity in the produced knowledge resource is a particularly sensitive issue. An early phase of the taste fabric construction process is to normalize the casually-stated keywords and phrases into formal ontologies of non-idiosyncratic form (e.g. “Nietzsche” → “Friedrich Nietzsche”, “dogs” appearing under the “passions” category → “Dog Lover”). Already, the unrestricted idiosyncratic language which bears traces of an authorship are beginning to be wiped away. In contrast, collaborative filtering systems maintain ratings for each user, and while users do not have to be named, even unnamed users are not anonymous, they are only pseudonymous. A user’s name is simply wiped away and replaced with a unique id (renamed from “John Smith” to “User #123”), but the profile’s integrity is intact. Because the number of instances is quite large in the space of tastes, it may be possible to recover the identities of pseudonymized because the constitution of profiles are quite unique. At the very least, maintaining any information structured around the notion of a user lends itself to the perception that privacy of the source data may be violated.

Rather than preserving individual profiles, the taste fabric simply uses these profiles to learn the strengths of connections on a network whose nodes already exist (they are simply an exhaustively enumeration of all features in the ontology). The method of the learning is non-linear so explicit frequency counts cannot easily be recovered. Thresholding and neighborhood definition are further lossful transformations which make details of the original application data virtually unrecoverable. The final structure is sanitary—it assures the anonymity of the data source, and is much easier to distribute.

Instance-based semantic webs and ethotic representation. In the Semantic Web community, ontology and metadata systems are often seen as top-down and bottom-up approaches to knowledge representation, respectively. To draw parallels with the artificial intelligence literature, ontology is a category-based representation, and metadata is a feature-based representation. However, taste fabrics introduces the notion of an *instance-based representation*, which we feel to be a promising methodology for the Semantic Web community that warrants further study, especially into the issue of scalability. An instance-based representation lacks categories or features, having only items and dense numerical connections between them. Knowledge is thus unpacked from the linguistic symbolism of a category or feature’s name, and instead, is found in *connectionism*—the flow of semantics through a graph of items. The shift from symbolic interpretation toward continuous interpretation parallels Zadeh’s efforts in attempting to soften the bivalence of logic representation by giving a *fuzzier*, more continuous account of meaning (Zadeh, 2004).

Instance-based representations are more appropriate for semantic recognition and semantic mediation because they offer continuous numerical interpretation of entity similarity. In taste fabrics, users, groups of users, and cultures can all be represented uniformly as clouds of node activations in the fabric. A taste fabric allows the meaning of a user’s keyword profile to be ‘relaxed’ into a semantic cloud which we term an *ethos*. Using ethotic representation, semantic mediation between two users or entities in the fabric can be computed quite easily as shared activation, and even effectively visualized. By interpreting an *ethos* as a membership into *k*-neighborhoods, the resource can be used to classify users or entities into an ontology of

neighborhoods (the organizing force of ontology, in fact, is still present in the resource via neighborhoods and identity descriptors). Instance-based representations and ethotic representations would be well-suited for semantic resources meant for mediation and classification in the Semantic Web.

6 Conclusion

This paper presented a theory and implementation of taste fabrics, a semantic mining approach to the modeling and computation of personal tastes for lifestyle, books, music, film, sports, foods, and television. Premised on philosophical and sociological theories of taste and identity, 100,000 social network profiles were mined, ontologically-sanitized, and a semantic fabric of taste was weaved. The taste fabric affords a highly flexible representation of a user in taste-space, enabling a keyword-based profile to be ‘relaxed’ into a spreading activation pattern on the taste fabric, which we termed a *taste ethos*. *Ethotic representation* makes possible many improvisational manipulations of meaning, for example, the taste-similarity of two people can be computed as the shared activation between two ethoi. Taste-based recommendation is already implied by a taste ethos, as all items within an ethos are intrinsically relevant to the taste of the individual. Indeed, an evaluation of recommendation using the taste fabric implementation shows that it compares favorably to classic collaborative filtering recommendation methods, and whereas collaborative filtering is an opaque mechanism, recommendation using taste fabrics can be effectively visualized, thus enhancing transparency and cultivating user trust.

Two models of taste-based recommendation—one quantitative based on shared activation, and one qualitative based on *k-nearest neighborhoods*—were presented. Recommendation, time and world-sensitive user representation, and interpersonal taste-similarity, were illustrated within a survey of three applications of taste fabrics.

This paper makes three contributions to the literature. First, it presents a novel mechanism for mining and modeling the taste-space of personal identities and interests. Second, the mining and weaving of taste fabrics from idiosyncratic social network profiles raises the issue of *sanitation* of knowledge resources, and this paper illustrated how ontology and non-linear correlation learning could be used to purge idiosyncrasy and prepare a general-purpose grade knowledge resource. Finally and third, in addition to ontology-based and metadata-based knowledge resources, taste fabrics introduces a novel third approach to the literature—instance-based fabrics, where the notion of ‘knowledge’ is a purely relational one. Fabrics, we suggest, excel at semantic mediation, contextualization, and classification, and may play a valuable role as a context mediator in a recently complicated Semantic Web of formal, semi-formal, and now, informal, entities.

Acknowledgement

This research was supported by a British Telecom Fellowship, an AOL Fellowship, and by the research consortia sponsors of the MIT Media Lab.

References

1. K. Aberer *et al.* (2004). Emergent semantics. *Proc. of 9th International Conference on Database Systems for Advanced Applications (DASFAA 2004)*, LNCS 2973, 25-38 Heidelberg.
2. Aristotle. (350 BCE). *Nichomachean Ethics*.
3. Danah Boyd. (2004). Friendster and publicly articulated social networks. *Conference on Human Factors and Computing Systems (CHI 2004)*. ACM Press.
4. S. Brin and L. Page. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.
5. K.W. Church, P. Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), pp. 22-29
6. A.M. Collins, and E.F. Loftus. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, pp. 407-428
7. Mihaly Csikszentmihalyi & Eugene Rochberg-Halton. (1981). *The Meaning of Things: Domestic Symbols and the Self*, Cambridge, UK: Cambridge University Press.
8. Paul Deane. (2005). A Nonparametric Method for Extraction of Candidate Phrasal Terms. *Proceedings of ACL'2005*.
9. C. Fellbaum (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.
10. Tim Finin, Li Ding, Lina Zhou and Anupam Joshi. (2005) Social Networking on the Semantic Web, *The Learning Organization: An International Journal* 12(5), 418-435. Emerald Group Publishing.
11. Erving Goffman. (1959). *The Presentation of Self in Everyday Life*. Garden City, NY: Doubleday.
12. R.M. Haralick, S.R. Sternberg, X. Zhuang. (1987). Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence* 9(4), 532-550.
13. J. Herlocker, J. Konstan and J. Riedl. (2000). Explaining Collaborative Filtering Recommendations. *Conference on Computer Supported Cooperative Work*, pp. 241-250
14. D. Jensen and J. Neville (2002). Data mining in social networks. *National Academy of Sciences Symposium on Dynamic Social Network Analysis*.
15. O.P. John (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66-100). New York: Guilford.
16. T.K. Landauer, P.W. Foltz, D. Laham. (1998). An introduction to Latent Semantic Analysis . *Discourse Processes*, 25, 259-284.
17. Vladimir Levenshtein (1965/1966). Binary codes capable of correcting deletions, insertions, and reversals, *Doklady Akademii Nauk SSSR*, 163(4):845-848, 1965 (Russian). English translation in *Soviet Physics Doklady*, 10(8):707-710.

18. Hugo Liu. (2003). Unpacking meaning from words: A context-centered approach to computational lexicon design. In Blackburn et al. (Eds.): *Modeling and Using Context, 4th International and Interdisciplinary Conference, CONTEXT 2003*, LNCS 2680 Springer, pp. 218-232.
19. Hugo Liu and Pattie Maes (2005a). InterestMap: Harvesting Social Network Profiles for Recommendations. *Proceedings of IUI Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research*, January 9, 2005, San Diego, CA, USA. pp. 54-59.
20. Hugo Liu and Push Singh. (2004). ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* 22(4). pp. 211-226. Kluwer Academic Publishers.
21. Pattie Maes *et al.* (2005). Ambient Semantics and Reach Media. *IEEE Pervasive Computing Magazine*. Submitted.
22. Andrew McCallum, Andres Corrada-Emmanuel and Xuerui Wang. (2005). Topic and Role Discovery in Social Networks. *Proceedings of 19th International Joint Conference on Artificial Intelligence*, pp.786-791.
23. Grant McCracken. (1991). *Culture and Consumption: New Approaches to the Symbolic Character of Consumer Goods and Activities*, Indiana University Press.
24. Grant McCracken. (1997). *Plenitude*. Toronto: Periph: Fluide.
25. M. Newman. (2001). Who is the best connected scientist? A study of scientific coauthorship networks. *Phys. Rev. E* 64.
26. B.M. Sarwar *et al.* (2001). Item-Based Collaborative Filtering Recommendation Algorithms. *10th Int'l World Wide Web Conference*, ACM Press, pp. 285-295.
27. Ferdinand de Saussure. (1915/1959). *Course in general linguistics* (W. Baskin, Trans.). New York: McGraw-Hill.
28. Friedrich Schleiermacher (1809/1998). General Hermeneutics. In A. Bowie (Ed.) *Schleiermacher: Hermeneutics and Criticism*, p. 227-268. Cambridge University Press
29. J. Serra. (1982). *Image Analysis and Mathematical Morphology*, London: Academic Press.
30. U. Shardanand and P. Maes. (1995). Social information filtering: Algorithms for automating 'word of mouth'. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 210-217.
31. Amit Sheth, Cartic Ramakrishnan & Christopher Thomas (2005). Semantics for the Semantic Web: The Implicit, the Formal and the Powerful. *International Journal on Semantic Web and Information Systems* 1(1), 1-18. Hershey, PA: Idea Publishing Group
32. George Simmel (1908/1971) How is society possible? In D. N. Levine (ed.) *On Individuality and Social Forms: Selected Writings*, University of Chicago Press, Chicago
33. S. Staab, S. Santini, F. Nack, L. Steels, and A. Maedche (2002) Emergent semantics. *IEEE Intelligent Systems*, 17(1):78--86.
34. Stanley Wasserman (1994) *Social Network Analysis: Methods and Applications*. Cambridge University Press

35. L. Wheeless and J. Grotz. (1977). The Measurement of Trust and Its Relationship to Self-disclosure. *Communication Research* 3(3), pp. 250-257.
36. Brian Whitman and Steve Lawrence (2002). Inferring Descriptions and Similarity for Music from Community Metadata. In "Voices of Nature," *Proceedings of the 2002 International Computer Music Conference*. pp 591-598.
37. Brian Whitman & Paris Smaragdis (2002). Combining Musical and Cultural Features for Intelligent Style Detection. *Proceedings of the 3rd International Conference on Music Information Retrieval*.
38. Lotfi A. Zadeh (2004). Precisiated natural language. *AI Magazine, Fall, 2004*. AAAI Press.