

Taste Fabrics and the Beauty of Homogeneity

Hugo Liu, Glorianna Davenport, and Pattie Maes

The Media Laboratory, Massachusetts Institute of Technology

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

Introduction

The quintessence of an individual's taste is her aesthetic sensibility and system of preferences. Online social network profiles, such as those appearing on Friendster and MySpace, are a veritable "show and tell" for taste—allowing individuals to perform *acts of taste* by declaring their favorite books, what music they love, and what their passions are. By mining these social network profiles *en masse* and analyzing how each taste instance (e.g. a book, an author, a band, a cuisine, etc.) is meaningfully correlated with every other, an underlying fabric of taste common across individuals can be inferred. This article theorizes and describes the mining of taste fabrics from social network profiles, and discusses the application of taste fabrics to recommender systems and to taste-based semantic mediation. We suggest that taste fabrics illuminates a new paradigm of "sense-maker" services for the Semantic Web—*homogenous connectionist networks* such as taste fabrics provide ways to situate, compare, and unify entities along semantic dimensions which are normally hard to quantify, such as the taste dimension. Fuller expositions on the topic of this article can be found in (Liu & Maes, 2005a; Liu, Maes & Davenport, forthcoming).

Favorite things as "taste acts"

An individual's taste is a unifying force beneath all of her favorite things. In gustatory taste, one's palette for food is systematic rather than arbitrary—she likes spicy foods, he likes Asian foods, they like gastronomy. Similarly, one's taste for music, books, and lifestyles is also systematic, but whereas food knows a few powerful ontologies capable of capturing the scope of the palette, such as basic flavors and ethnic cuisines, taste for things in general is far harder to ontologize because its systematicity is quite latent and no few categorizations capture its richness with any satisfaction. But just because something eludes ontology does not mean that it is not precise or that it is uncapturable. Within the context of a social network profile, each favorite thing that an individual lists is a reified instance of their taste. While in isolation, each instance has ambiguous meaning (there are many reasons why someone might like Madonna), the instances taken all together demonstrate more coherence and thus should illuminate something very specific—that individual's cultural taste.

The idea that someone telling you their favorite things can help you (or a machine) to infer their taste is well explored in Speech Acts Theory. John Austin's formulation of the theory (1962) distinguished between surface utterances and underlying acts. When your boss says "there's no room in this company for mediocrity," the surface utterance seems to state a moral truth that the company is not mediocre, but the *act* or intent of the utterance can mean to make a threat. Likewise for taste, an individual listing "Marilyn Manson's Antichrist Superstar" as a favorite album (Manson is a goth rock musician) and "Salinger's Catcher in the Rye" as a favorite book constitutes two factual utterances, but the intent and spirit, or *illocutionary force*, underlying the

two utterances seems contemptuous of social and ethical norms, and so the illocutionary act here, is potentially one of rebellion. We suggest that a social network profile, as an enunciation of one's favorite things, can be viewed as an impressionistic portrait of an individual's taste—with a sense of totality of taste efflorescing from the intersection of all of these taste acts.

To make sense of an individual's favorite things, we need to discover what taste act each instance is trying to communicate, and then we need to infer the gestalt of the individual's taste from the intersection of all these acts. One way to model this is to take a page from lexicography—be reconciled to the fact that taste is hard to ontologize and resort to ethnography—look for patterns of co-occurrence within the space of taste instances. Mine all profiles, and for each profile extract its instances and record the pairwise associations between these instances. Proceeding as such over all profiles, some pairs of instances will develop strong affinities, while the co-occurrence of other pairs of instances will be revealed as statistically arbitrary. We call this resultant matrix of pairwise associations a *taste fabric*. On a taste fabric, each instance has a corresponding node. All other nodes strongly linked to this node, plus all the two-hop nodes linked strongly to the one-hop nodes, constitute its *space of possible taste acts*, and this space is defined by spreading activation (Collins & Loftus, 1975) outward from the origin node. If only one instance is given, the space of taste acts is ambiguous, but if two instances of an individual's favorite things are given, their spaces of taste acts intersect. As the number of given instances increases, the shared taste act space is increasingly constrained and clarity emerges out of ambiguity. For a typical social network profile of twenty or so instances, after spreading activation outward from each instance's node, a shared activation pattern will emerge. We call this pattern the taste characteristic or *taste ethos* of the individual. In summary, we represent an individual's taste by the taste ethos consequent from his social network profile. The computation of taste ethoi is only possible if we can mine the taste fabric underlying social network profiles.

Mining taste fabrics

In early 2004, we mined 100,000 social network profiles from two online social networks. By applying lightweight natural language processing, ontology-driven normalization, and statistical language modeling to the text of these profiles, we constructed a taste fabric and devised methods for computing taste using the fabric.

Normalization. The social network profiles we examined are broken down first into categories such as “general interests,” “favorite books,” “favorite sports,” “favorite foods,” “favorite music”, and within each category are instances, usually given as a punctuation-delimited list. After the plaintexts of profiles are downloaded, the intra-category texts are segmented using regular expressions, with 75% success for “general interests” and 90% success for all other categories. After segmentation produces lists of idiosyncratic natural language fragments, a formal ontology constructed from various sources such as imdb.com (for film metadata), allrecipes.com (for food metadata), allmusic.com (for music metadata), and wikipedia.org (for “identity” metadata) is used to *clean* idiosyncrasy from these fragments. Indicative examples of this normalization: “Lynch's Twin Peaks” into {film-name: ‘Twin Peaks’, film-auteur: ‘David Lynch’}; “Bach” into “J.S. Bach” rather than into “C.P.E. Bach,” using popularity metadata to assist this disambiguation; and common misspellings like “Bahc” into “Bach” using Levenshtein edit distance (1965). Furthermore, since our ontology has metadata links between film directors and

their films, music artists and their albums, etc., idiosyncratic fragments are mapped into both the cleaned fragment, but also into its related metadata using a 0.5 discount factor, e.g., “twin peaks” into (“Twin Peaks”, weight=1.0) and into (“David Lynch”, weight=0.5). This way, the recognition of favorite things can be said to be *situated* within its semantic context.

Weaving. Post-normalization, profiles are composed of a bag of descriptors containing two kinds—*identity descriptors* from the “general interests” category, e.g., “intellectuals,” “fashionistas,” “dog lovers,” and *interest descriptors*, from all other categories, e.g. favorite books, music, foods, etc. The key premise for mining taste is that an individual’s taste is self-coherent and expressive of a latent sensibility—and from this, it follows that there should be mutual information between all of one individual’s descriptors. So to weave the taste fabric, a machine-learning algorithm looks at each profile, and records all the pairwise associations between its descriptors. Using a common statistical language modeling metric known as *pointwise mutual information* (PMI) (Church & Hanks, 1990), the completed taste fabric looks like an n by n correlation matrix giving the numerical affinity between any two descriptors. After stripping away null rows, the resultant raw taste fabric is a 12,000 by 12,000 matrix (600 of the 12,000 descriptors indicate identity). A final step of post-processing carves up the raw taste fabric into various *taste neighborhoods* by applying mathematical morphology (Haralick, Sternberg & Zhuang, 1987) in conjunction with spreading activation (Collins & Loftus, 1975).

Computing taste

Taste fabrics can be used to perform *situated recognition* of a person or persons from a keyword-based profile, to make taste similarity measurements between entities, and to produce taste-based recommendations.

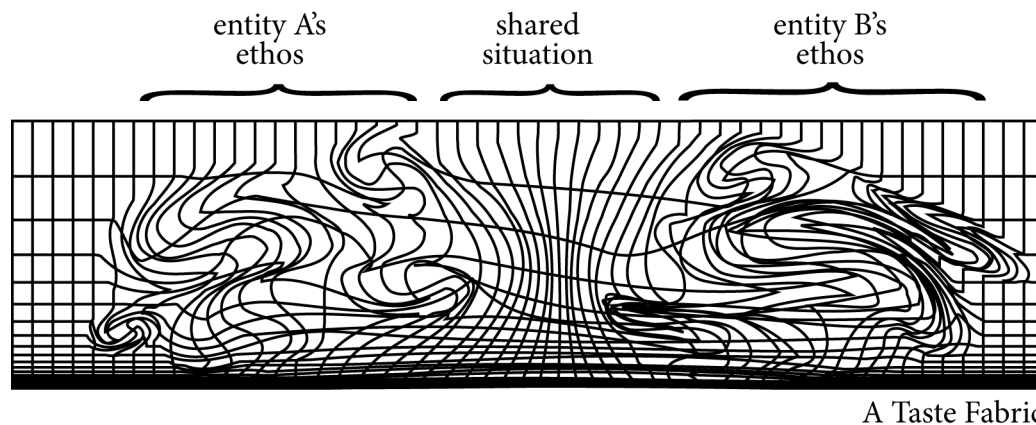


Figure 1. On a taste fabric, a person, group, or culture is represented by a unique network activation pattern known as a taste ethos. Where two activation patterns intersect defines their shared taste situation.

Situated recognition. As depicted in Figure 1, a person, group, or culture has an *ethotic representation* in the taste fabric—that is to say, an entity is not merely a discrete list of interest keywords, but rather it is a *location* won the cultural space of taste, and is represented as a unique activation pattern in the fabric, called a *taste ethos*. To situate a person, group, or culture into the taste fabric, the entity’s textual profile of favorite books, music, etc. is normalized in the same manner as in the mining phase such that post-normalization, the entity’s profile is a bag of

(weighted) normal descriptors. Treating the taste fabric as a spreading activation network, the nodes corresponding to the entity's profile are activated, and activation is spreading outward from these seed nodes with some discount d —for example, a smaller d is a more lenient discount which begets a broader ethos. Compared with a discrete list representation, ethotic representation is semantically more flexible, and because a fabric is densely connected, even almost continuous, entities can be defined and manipulated with high resolution. Additionally, ethotic representation is well-suited for visualization—in InterestMap (Liu & Maes, 2005a), the taste fabric is depicted graphically as a spring loaded semantic network with visual features such as identity hub-and-spokes, taste cliques, and taste neighborhoods. Taste ethoi are visualized atop the InterestMap by using gradient colors to highlight the activated region in the network. IdentityMirror (Liu, Maes & Davenport, forthcoming) is an application that uses taste fabrics to visualize a “reflection” of an individual as a keyword cloud of his taste ethos. To pose self-reflection as being situated with the cultural zeitgeist and influenced by ephemera such as mood, the mirror shows different facets of an individual, produced by biasing the ethos using topics and trends extracted from daily news feeds, and from an individual's weblog and playlists.

Taste similarity measurement. The taste similarity between two or more entities can be computed as their shared situation—the weighted nodal intersection of the individual activation patterns. This yields a quantitative measurement that would need to be interpreted further. For example, quantification could advise that John is closest in taste to Mary out of a hundred people, but just looking out-of-context at the numeric score of the affinities between John and Mary does not yield much information. Furthermore, the delineation of *taste neighborhoods*—that is to say, larger patches of nodes of the fabric organized around the thematic of identities—affords qualitative taste similarity measurements. In this case, each entity's ethos is mapped in k nearest neighborhoods that the entity “inhabits,” and shared situation is computed as the list of neighborhoods inhabited by both entities. This qualitative measurement is more appropriate for applications that wish to verbalize explanations for the computed similarity between individuals, as taste neighborhoods are well described by their identity thematics. Ambient Semantics, *cf.* (Liu, Maes & Davenport, forthcoming) is an application which uses taste fabrics to facilitate social introduction between strangers by illustrating shared tastes using the methods described here.

Recommendation. Recommendations can be made in the same manner as taste ethoi are created. Since the fabric is structured according to taste affinities, spreading activation outward from a seed profile will continuously harvest nodes suitable for recommendation. The stronger a node's activation is, the more proximal that recommendation should be to an individual's taste. However, proximal nodes given as recommendations also run the risk of being too obvious or already known to the individual. Just as IdentityMirror biased an ethos to highlight particular facets of an individual's taste, so recommendation can also be organized according to particular facets of taste by introducing a contextual bias in the spreading activation, *e.g.* “show me things I might like, along the lines of what Jane likes.” Evaluation of taste fabrics in a graded recommendation task showed that it performed slightly better than collaborative filtering, *cf.* (Liu & Maes, 2005a)—while affording the added benefit of being easier to visualize and explain, which might improve the trust quotient of the recommender.

Sense-making for the Semantic Web

As the Semantic Web proceeds to bring semi-formal and informal entities into its fold, it will need more powerful ways to mediate their varying semantics. We suggest that taste fabrics indicate a broader class of *semantic fabrics* that could serve as context mediators and sense-maker services for a more improvisational semantic web. For example, to generate Semantic Web entities from Internet Newsgroups, our methodology could be applied to create semantic fabrics for various dimensions relevant to newsgroups, such as opinions or user demographics, then these fabrics could be used to situate, compare, and unify newsgroups, producing metadata and ontology that will be useful to any formal representation for entities.

Return to homogeneity. A key differentiator of semantic fabrics from traditional metadata or ontology classification systems is that fabrics produce measurements of higher semantic resolution. Traditional metadata and ontology systems are prone to the *inferential distance problem*—for example, in WordNet (Fellbaum, 1998), a lexical ontology, it takes seven link-hops to discover that a cat is an animal (*i.e.*, cat is-a1 feline is-a2 carnivore is-a3 placental is-a4 mammal is-a5 vertebrate is-a6 chordate is-a7 animal). This is not even mentioning that the commonsensical “cat is a pet” is not even in that knowledgebase. In contrast, fabrics are densely, if not fully connected, so spreading only one or two link-hops outward from a node will uncover its full semantic context. Fabrics also buck a trend toward heterogeneity in semantic representation. Whereas ontology and metadata systems are bloating with semantic relation diversity, a semantic fabric is homogenous in its relation type (they are all affinity scores) and its method of reasoning (spreading activation). Because a semantic fabric represents just one semantic dimension, it can do it exhaustively and yield a very high-resolution resource for making semantic measurements along that dimension. Semantic fabrics re-introduce continuity into an increasingly symbolic and discretized semantic universe, begging us to heed that old adage, “do one thing, and do it well.”

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. John L Austin (1962). *How to Do Things with Words*. Cambridge (MA): Harvard UP.
2. K.W. Church, P. Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), pp. 22-29
3. A.M. Collins, and E.F. Loftus. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, pp. 407-428
4. C. Fellbaum (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.
5. 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.

6. 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.
7. 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.
8. Hugo Liu, Pattie Maes and Glorianna Davenport (forthcoming). Unraveling the Taste Fabric of Social Networks. *International Journal on Semantic Web and Information Systems*.