Shopping for Top Forums: Discovering Online Discussion for Product Research

Jonathan L. Elsas*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jelsas@cs.cmu.edu

Natalie Glance
Google, Inc.
Pittsburgh, PA 15213
nglance@google.com

ABSTRACT

Community generated content, or social media, has become increasingly important over the past several years. Social media sites such as blogs, twitter and online discussion boards have been recognized as valuable sources of market intelligence for companies wishing to keep abreast of their customers' attitudes expressed online. There has been little focus, however, on providing a similar service to potential customers.

In this paper we present a system for aiding consumers with their product research by providing access to community generated content. We focus specifically on online forums or message boards, which are particularly useful for product research. These web sites often host discussion among users with first-hand product experiences, expert users and enthusiasts.

The system presented here is designed to integrate with a shopping search portal, providing access to online forums that are likely to have a significant amount of discussion relating to a user's expressed interest in product brands and categories. We describe this system and present experiments showing that in the context of a shopping search engine, the proposed system is preferred or equivalent to results from a web search engine 80% of the time and achieves accuracy at the top ranked result of 85%.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Online discussion forums, message boards, product search

*This research was done while at Google.

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

1. INTRODUCTION

Consumers researching products for the purposes of making purchasing decisions frequently visit online shopping portals sites. These sites such as Google Product Search[1], Bing Shopping[2] or Yahoo! Shopping[3] aggregate many types of content for the consumer: editorial and user reviews, buying guides, and price comparison tools. But missing from the current product research landscape is the presence of large-scale conversational reviews, such as those found on online forums and discussion boards. In these sites, frequently many authors share their first-hand experiences with products, as well as troubleshooting tips, advice, or general discussion.

There are an enormous variety of online forums on the web, generally topically focused and often cultivating active communities of enthusiastic contributors. These types of social media outlets, however, can be difficult to discover by individuals who may not already be familiar with the community. The current tools to access online forum archives are lacking, and although web search engines index online forum data, many distinguishing characteristics of online forums are ignored by traditional ad-hoc information retrieval techniques. Additionally, to our knowledge, there are no publicly available tools to help in identifying forums, rather than forum threads or posts.

This paper addresses the task of identifying discussion forums rich with product-related discussion. In these forums a potential shopper may find first-hand reviews, product comparisons or other user experiences. We approach this task as an information retrieval problem, ranking forums with respect to product search related information needs. This system is designed to integrate with a shopping portal to provide users with access to archives of community generated commentary as well as a forum to interact with experts and enthusiasts when making purchasing decisions.

The main contribution of this work is on a novel forum ranking model (Section 4.3), aimed at identifying online forums with a high density of discussions on product-related topics. This ranking model leverages a rich set of document annotations: document classifications, identification of the structure within the forum, annotation of product mentions, and categorization of those mentions to a product ontology. The ranking model achieves greater than 85% precision at the top ranked result and is preferred or equivalent to web results restricted to online forum pages 80% of the time.

[1] `http://www.google.com/products`
[2] `http://www.bing.com/shopping`
[3] `http://shopping.yahoo.com/`

2. MOTIVATION AND TASK DESCRIPTION

There are three main use cases for online shopping: product navigation, browsing and product research. A complete shopping experience must support all three to be successful as a destination site. Shoppers doing research prior to making a purchase tap into many kinds of online information, in particular they may seek out editorial or user reviews of specific products, buying guides for categories of products or informal conversational product discussion such as those found in message boards.

Message boards, or discussion forums, are an especially good place to find product comparisons within a category of items, to find expert opinions, and to find first-hand product experiences. But, these outlets are rarely integrated into online shopping sites. Some shopping sites address this by creating their own set of forums, but these are not necessarily successful at attracting the critical mass of expertise to be useful for aiding shopping decisions. In many cases, there already exists incredibly rich message boards with well established communities and large archives of product-related discussion. These message boards may be run by product manufacturers (such as discussions.apple.com), brand enthusiasts (such as forums.macrumors.com), independent interest groups (such as dpreview.com or gpsreview.net) or professional reviewing organizations (such as forums.cnet.com).

We propose to tap into these existing rich outlets of product discussion by pulling online forum results from the web into the user interaction flow of the shopping site. In order to do this, we must address the questions: when do we choose to show discussion forum results and what exactly do we show?

2.1 Information needs in shopping portals

There are may ways shoppers may express their information needs in a shopping search portal, for example by typing a search query into a search box or by clicking on product facet values to restrict the results show. Let's consider the first question about when to trigger forum results when a query is entered to a search box. What kind of query falls into the product research bucket? Searches for particular items or for one of a product line, like "HP Laserjet 1020", can be interpreted as product navigation queries. In this case the user's intent to find information about a particular product is clear, and the best result is to provide pricing information and reviews for products that match the query. Searches for a broad category of product, like "microwave oven," may be interpreted as seeking a browsing entry-point for that category. In this case, we can argue that the user is best served by being shown, for example, a set of top brands, best-selling products and buying guides, or other tools to narrow down the product landscape.

The third bucket of product research queries fall between the specific navigational query and broad product category queries. There are queries like "Bose speaker" or "Apple laptop" where the user has some notion of a relevant category and brand, but has not yet narrowed down to a specific product or product line. While the existing product portal content such as buying guides, product reviews and price comparisons are still likely to be helpful, consumers may also be interested in more informal sources of information. Online forum discussions can serve this purpose, as rich sources of product comparison information, product support, trouble-shooting and informal reviews on a range of items in the same class. Perhaps equally important, providing access to the right forums not only provides content likely to be useful, but also provides access to experts and enthusiasts willing to answer questions.

Thus, we frame our task to be: finding top forums relevant to this third bucket of queries, characterized roughly as brand-category pairs. Although we discuss possibly identifying these types of queries from the text entered in a search box, there are other ways for a user to express their interest in a in a category-brand pair. For example, a user may select fact values in a browsing interface in order to limit the displayed items, as in Figure 1. Or, we may be able to identify

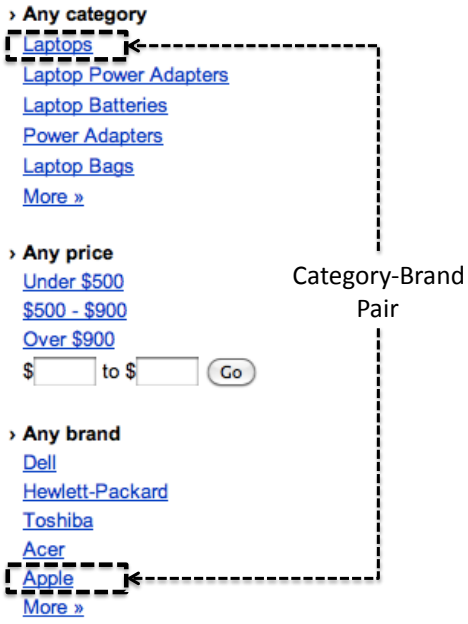

Figure 1: Example facet value selection in a product browsing interface, with the category-brand pair query `(Laptops, Apple)` selected.

the shopper's intent implicitly through recent interactions with the system. We leave as an orthogonal task the job of identifying such expressions of a user's information either from the query stream of a search engine or other means. For the remainder of the paper, we will assume a user has expressed an information need in the form of a category-brand pair, which we will refer to as the *query*.

2.2 Addressing Category-Brand Queries

The second main question is what consitutes the search result, and especially, what is the correct level of granularity for the search result? Online forums are typically organized hierarchically: an online forum site often has several high-level topical forum categories, which are split into finer-grained categories. Each of these contains many threads, collections of user-contributed messages. Should the forum results be top-level forum site, a lower-level forum, a message thread, or message? Top-level forums are almost always too broad and topically diverse to be useful as a result. Sending a user to a top-level forum generally means the user still needs to search within the forum. Returning posts is unhelpful in the opposite extreme: taken out of context, an individual post is rarely informative. The sweet spot over-

all seems to be returning both the forum, plus a list of the most relevant threads. As with most web search results display, we choose to provide not just the the online forum, but also contextual information with the result. In this case, the contextual information includes the top ranking thread titles with metadata possibly including the number of messages or a date range of posting times.

Figure 2 shows an example organization from an online forum. In this example, we can see the hierarchical organi-

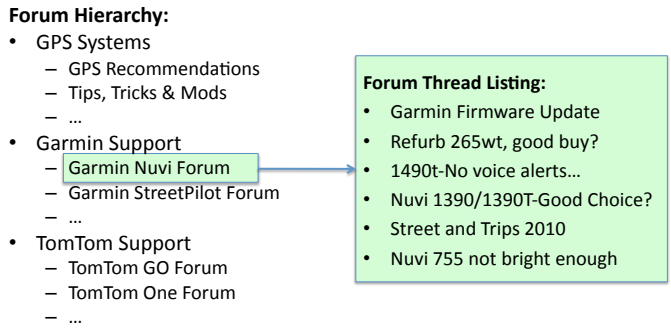

Figure 2: Example hierarchical organization of an online forum, gpsreview.net. Topical forum organization shown at left, and thread listing in a single forum shown at right.

zation of an online forum on the left, with a thread listing on the right. Given a shopper's interest in a category-brand pair, such as `(GPS, Garmin)`, we may deem the forum entitled "Garmin Nuvi Forum" a relevant result. In this case, any individual thread within this forum would be too specific, and the forum site as a whole would be too general. In general, the leaf-node forum may not always be the best choice of a result. For example, a higher level in the hierarchy (eg. the "Garmin Support" forum) may be a better choice in this case. But, the leaf-node forum tends to provide a good trade-off between the generality of the forum site and specificity of a message or message thread. We leave for future work investigation of identifying per-query the appropriate level of hierarchical organization.

3. RELATED WORK

Social media, including online discussion forums, have been the focus of much recent research. In the area of information retrieval, the TREC blog track has focused the research community on techniques for ranking and opinion mining of blogs and blog posts with respect to user queries [5, 8]. Recently several models for thread retrieval in online message boards have been proposed [3, 11]. This previous work on retrieval in online forums has focused on the *message thread* as the primary unit of retrieval, whereas in this work we are concerned with ranking *forums*, or collections of threads. The forum ranking model presented below builds upon this previous work studying blog retrieval and message thread retrieval [1, 3].

Online forums have also been the focus of several data mining studies. Wanas et al. [12] developed methods to automatically identifying high quality quality posts in a large discussion board. Yang et al. [13] apply information extraction and techniques to the task of automatically identifying online forum structure from web pages, such as segmenting threads into messages, identifying author names and message posting dates. Zhang et al. [15] present a study of the social dynamics in online forums to identify author expertise.

Online forums and blogs have been recognized as fertile ground for mining product discussion. Glance et al. [4] present a system for mining online discussion for the purposes of monitoring popular opinion about brands or products. This system provides facilities for extracting threading structure from online discussion boards, opinion mining and aggregation, and social network analysis. The work presented here similarly focuses on finding product discussion in online forums, but for the goal of aiding consumers in their product research rather than aiding companies monitor popular opinion.

4. FORUM RANKING APPROACH

Our approach to identifying forums with rich product discussion is based on two levels of information aggregation:

- From lower-level product mentions to higher level product brands and categories.
- From lower-level messages to collections of messages and threads, the forums.

Both of these aggregation steps require rich levels of document annotation, as well as a model for scoring forums to aggregate from the message level.

The focus of this work is on the forum ranking model (Section 4.3) but we provide a high-level description of the annotations used in the following sections. There are numerous automatic techniques for document classification, structure extraction, product annotation, and product name normalization [9, 10, 13], and the details of those applied here are out of the scope of the current work.

4.1 Product Annotation

In each document in our collection, all references to products, product lines, and brands are annotated. Each annotated product mention is mapped to a single entry in a product catalog. This catalog contains all known brands, product lines and products, and each entry in the catalog is associated with one node in a product category ontology. This ontology providing a hierarchical organization of products, useful for faceted search and browsing in the product search portal.

An illustration of the product annotation and mapping to nodes in the product category ontology is shown in Figure 3. In this figure, we can see a span of text containing two product mentions, one to a brand ("Switcheasy") and one to a product line ("Switcheasy Vulcan"). Both of these spans of text are annotated as product mentions and assigned a mapping to a node in a product catalog. Note that neither of these mentions refer to the specific product. Each entry in the product catalog is mapped to a node in the product category ontology, in this case the "MP3 Player Cases" leaf node. The resulting annotations in the text correspond to the category-brand pair `(MP3 Player Cases, Switcheasy)`.

4.2 Forum Structural Annotation

In addition to the product annotation, we also produce annotations of the online forum structure in our collection.

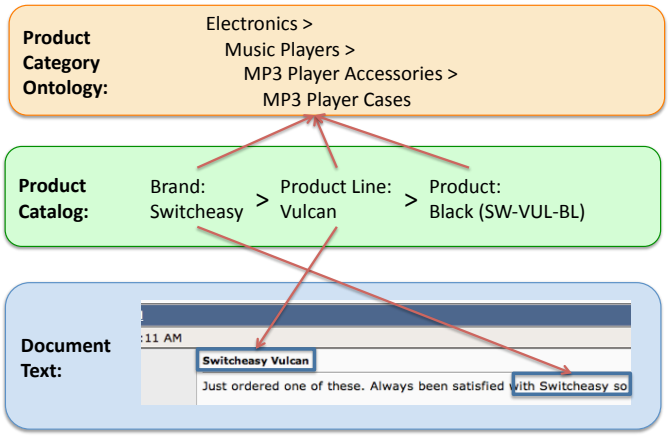

Figure 3: An example annotation and mapping from product mentions in a document text (bottom) to nodes in the product category ontology (top).

See Figure 4 (dashed lines) for example extracted structure. The following fields and attributes are extracted from these documents:

- **Message-Level Annotations**: post date, author name, body text
- **Thread-Level Annotations**: parent forum, number of messages, title text

Each message thread is assigned to a single *parent forum* containing that thread. Online forum sites are typically organized hierarchically, containing many forums within a single site. We make the assumption that each message thread belongs to only the immediate parent forum in that site, typically a leaf node in the forum hierarchy. In the example showin in Figure 4 (top), the thread shown belongs to the "iPhone Accessories" forum, and other higher-level forums (eg. "iPhone Forums") are ignored.

4.3 Forum Ranking Model

We approach the task of identifying forums with rich product discussion as an information retrieval problem, ranking forums with respect to a category-brand query. We take a probabilistic language modeling approach to scoring the online forums with respect to the query. In our approach, we aggregate information from the lowest-level in the forum hierarchy, the message text, to the level we're interested in scoring, the forum. We rank forums by their conditional likelihood given the query, $P(f|q)$. The estimation of this probability is shown below, first applying Bayes theorem and marginalizing over the message threads in the collection t. Letting f be the forum, and q be the user's query:

$$\text{Score}(f,q) = P(f|q) \overset{\text{rank}}{=} P(f) \sum_t P(t|f) P(q|t). \quad (1)$$

Note, we drop the probability of observing the query $P(q)$ as it is doesn't influence the thread ranking when the query is fixed. We assume a uniform prior probability of forum relevance, $P(f)$, and the probability $P(t|f)$ is given by

$$P(t|f) = \begin{cases} \frac{1}{|f|+\alpha_f} & \text{if thread } t \text{ belongs to forum } f \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $|f|$ is the number of threads in forum f and $\alpha_f \geq 0$ is a discount penalizing forums with few thread.

In keeping with previous research on ranking structured documents [7], we model the query likelihood with respect to the thread, $P(q|t)$, as a mixture model. The mixture components in this case, are the thread title language model θ_t and the message body language models θ_m. Marginalizing over the messages in the collection, we have

$$P(q|t) = \lambda P(q|\theta_t) + (1-\lambda) \sum_m P(m|t) P(q|\theta_m) \quad (3)$$

and we similarly define $P(m|t)$ as

$$P(m|t) = \begin{cases} \frac{1}{|t|+\alpha_t} & \text{if message } m \text{ belongs to thread } t \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $|t|$ is the number of messages in thread t and $\alpha_t \geq 0$ is a discount penalizing threads with few messages. Note that the query likelihood given the thread $P(q|t)$, in addition to being an additive component of the forum scoring (Equation 1) also provides a natural means to identify top relevant threads from within a forum.

The normalizing terms above (Equations 2 and 4), provide parameters to discount the score for threads with few messages and forums with few threads. We can interpret these probabilities as assuming there are some number of "unseen" threads and messages (α_t and α_m) in the collection with a zero score. Similar normalization techniques have been shown to be effective in blog feed search [1].

The language model probabilities above ($P(q|\theta_t)$ and $P(q|\theta_m)$ in Equation 3) are estimated as multinomials with Bayesian smoothing using Dirichlet priors [16]. In this setting, we are not scoring documents with respect to the degree of *textual match* with the query, but rather to favor documents with a high density of product category mentions. To this end, we treat a product mention as equivalent to a text token in the document.[4]

When scoring against the thread title language model, we calculate these probabilities as:

$$P(q|\theta_t) = \frac{n(q,t) + \mu_t P(q)}{|t| + \mu_t} \quad (5)$$

where $n(q,t)$ is the the the number of times the query q was mentioned the title, $|t|$ is the number of text tokens in the title, and μ_t is a smoothing parameter. When scoring against the post text, we have two background language models: the collection (as above) and the entire thread body. To take advantage of these two background models, we apply two-stage Dirichlet smoothing [14], giving the following:

$$P(q|\theta_m) = \frac{n(q,m) + \mu_m \frac{n(q,T)+\mu_T P(q)}{|T|+\mu_T}}{|m| + \mu_m} \quad (6)$$

where $n(q,m)$ and $n(q,T)$ is the the the number of times the query q was mentioned the message body and thread body respectively, $|m|$ and $|T|$ are the number of text tokens in the message and thread body, and μ_m and μ_T are smoothing parameters. In all the above, $P(q)$ is the *background probability* of observing a mention of the category-brand pair in the collection, estimated via maximum likelihood.

[4]This treatment of a product mention as a single text token is identical to the scoring used by the Indri information retrieval engine (`http://lemurproject.org/indri`) when scoring annotated text with the `#any` operator.

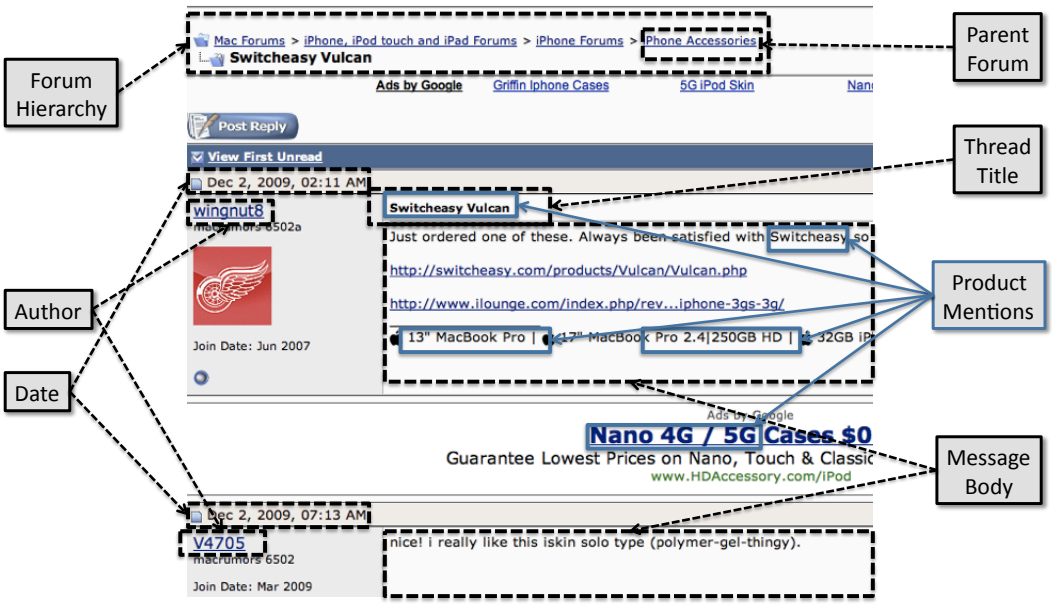

Figure 4: An example forum thread page (from forums.macrumors.com), showing forum structural annotations (black, dashed lines) and product mention annotations (blue, solid lines).

The ranking model here is derived for a single category-brand pair query. This is the equivalent to a single-term query in a text search engine, and avoids the difficulty in combining relevance scores across several terms which may have different collection-level characteristics or significance to the underlying information need. This model, can easily be extended to multi-term queries in the same way that standard language modeling retrieval models approach the problem [6]. Because we focus on single-term queries in this work, the model is not very sensitive to the language modeling estimation details (ie. calculation of $P(q|\theta_x)$). If the model is extended to handle multi-term queries, closer attention must be paid to these estimation details.

5. DATASET DESCRIPTION

The dataset used in the following study is a richly annotated document collection, as well as a product catalog and product category ontology. The documents used are a sample of webpages from online forums from the first tier of a commercial search engine index, excluding those documents identified as pornography and spam. The forum structure is extracted, as described above in Section 4.2. The document text is annotated with product mentions and those mentions are mapped into a product category ontology as described in Section 4.1. For the purposes of this study, we focus only on consumer electronics products.

The final dataset contains over 3.5 million online forums, with almost 400 million messages organized into over 40 million message threads and contributed by over 45 million authors. Almost 40% of the message threads containing at least one product mention, and there are over 350 million total mentions in the collection, corresponding to 95 million unique category-brand pairs.

6. EXPERIMENTAL SETUP & RESULTS

We evaluate the system presented here along two dimensions. In this section, we refer to the forum ranking produced by the above algorithm as the *Top Forums* results. First we look at a precision-oriented evaluation of the top results ranked by this system. Because of the limited screen real estate in a product search portal, it is likely that only a small number of online forums can be presented to a user. For this reason, we view precision at the top ranked results as the primary measure of performance for this task.

Second, we compare the ranking produced by the system to a commercial search engine ranking, limiting those results to only pages from online forums. The search engine queries used are a simple concatenation of the category name and brand, for example `[garmin gps]` or `[hewlett packard laptops]`. Both evaluations use a set of 96 queries sampled to represent both popular and unpopular brands and categories, restricted to "concrete" categories roughly corresponding to leaf nodes in our product category ontology.

The evaluation dataset was collected through a web interface, shown in Figure 5. Participants were shown a list of category-brand pairs, and asked to select one from the list. After doing so, the top ten results from the system described here were presented, and participants were asked to identify those results deemed relevant to the query. Participants were also shown the top ten results from a commercial search engine, restricted to pages only from online forum websites, and asked to identify which results they would find more helpful in making purchasing decisions or performing product research.

Figure 6 shows example results for several queries. These results show the top 3-4 results for the queries `(Headphones, Sennheiser)` and `(GPS, Garmin)`, as well as the top-scoring 2-3 threads in each forum. Each result provides the forum and thread titles, and some metadata indicating the sizes of these threads and forums. The displayed threads are those that have the highest contribution to the forum score, their score given by Equation 3. These results clearly demonstrate

the model's ability to identify threads with a high density of product mentions. In these results, the retrieved threads are direct product comparisons, reviews, purchasing advice and other related content.

6.1 Algorithm Parameter Tuning

The ranking algorithm above has several parameters that can be tuned to change the characteristics of the system output. Table 1 shows the parameter values used for this evaluation. These parameters were identified through manual tuning prior to the evaluation. We leave for future work investigation of automated methods for tuning parameters in this retrieval model.

Parameter	Description	Value
α_f	Small forum discount	200
α_t	Small thread discount	50
λ	Title weight	0.8
μ_t	Thread title smoothing	300
μ_m	Message body smoothing	1000
μ_T	Thread body smoothing	2500

Table 1: Parameter settings used in the experiments..

6.2 Precision-Oriented Evaluation

As a component to a product search and browsing portal, where screen real-estate is is shared with browsing controls, product listings, merchant offerings and advertisements, a high-precision ranking is crucial. For this reason, we focus on the precision of the ranked list of top forums at the first five rank positions. Table 2 shows the precision of the system averaged across all 96 queries. From this figure, we can see that precision at the top rank exceeds 85%, and drops to just under 80% at rank position five.

Figure 3 shows the precision at the top rank position for the product categories with the most queries in our dataset. From this figure we can see that there is a range in performance across categories, with some categories like "handhelds & pdas" retrieving a relevant result at the top rank 100% of the time. Other categories like "computer monitors" have much lower precision, only retrieving a relevant forum 63% of the time.

6.3 Side-by-side Evaluation

In addition to the precision oriented evaluation described above, we also evaluated the Top Forums results against those of a web search engine restricted to online forums pages. Participants were asked to identify the most useful set of results for product research in the context of an online shopping experience. They were given three choices: Top Forums results, web results restricted to online forum pages, or neither.

Rank Cutoff	Precision
1	0.8511
2	0.8191
3	0.7943
4	0.7926
5	0.7894

Table 2: Precision at top five rank cutoffs, averaged across all 96 queries.

Category	Num. Queries	Precision @1	@2	@3
computer monitors	8	0.625	0.563	0.417
home theater systems	9	0.667	0.778	0.741
radios	8	0.750	0.750	0.750
memory	8	0.875	0.857	0.810
digital cameras	11	0.909	0.864	0.849
bridges & routers	5	1.000	1.000	0.933
handhelds & pdas	7	1.000	1.000	1.000
digital cameras	10	1.000	1.000	1.000
flat panel televisions	10	1.000	0.950	0.933

Table 3: Precision at top rank per category.

Table 4 presents the results of that study, showing that the top forums results are as good or better than the web results 83% of the time.

System	# Queries Preferred
Top Forums	46 (48%)
Neither	34 (35%)
Web Results	16 (17%)

Table 4: Annotator preference for Top Forums results or filtered Web results.

7. CONCLUSIONS & FUTURE WORK

This paper presents a system for identifying online forums rich in product discussion relating to category-brand pairs. The algorithms proposed here perform two levels of aggregation to find forums relevant to these higher-level concepts. First, using a product catalog and category ontology, we aggregate specific product mentions to category-brand references. Second, using structural annotation of the online forum, we aggregate relevance scores from online forum messages and threads to forum scores. The system achieves over 85% accuracy in identifying the top-ranked online forum result, and is preferred or equivalent to web search results 83% of the time.

Although the system presented here focuses on identifying forums relevant to category-brand pairs, the approach is much more general. For example, with the same annotation scheme, forums could be scored with respect to category or brand only. Or, as mentioned above, the equivalent to multi-term queries could be run in the system to identify forums discussing several brands in the same category.

The algorithm presented here identified top forums based on the density of *discussion threads* relevant to the query. An alternative approach would be to identify top forums based on the density of *expert authors* with respect to the topic of the query. A fertile area for exploration is the application of expert finding models [2] to this task, and the combing of thread- and expert-models to finding top forums.

Finally, these aggregation techniques could also be applied to improving general web search, specifically when scoring online forum content. Currently published models for web search do not take into account the unique structure of on-

line discussion boards. Leveraging that structure within web retrieval models may yield further improvements.

8. REFERENCES

[1] J. Arguello, J. L. Elsas, J. Callan, and J. G. Carbonell. Document representation and query expansion models for blog recommendation. In *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM New York, NY, USA, 2006.

[3] J. L. Elsas and J. G. Carbonell. It pays to be picky: an evaluation of thread retrieval in online forums. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 714–715, New York, NY, USA, 2009. ACM.

[4] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428, New York, NY, USA, 2005. ACM.

[5] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 blog track. In *Proceedings of the 2007 Text Retrieval Conference*, 2007.

[6] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40(5):735–750, 2004.

[7] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 143–150, New York, NY, USA, 2003. ACM.

[8] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 blog track. In *Proceedings of the 2008 Text Retrieval Conference*, 2008.

[9] S. Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, 2008.

[10] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[11] J. Seo, W. B. Croft, and D. A. Smith. Online community search using thread structure. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1907–1910, New York, NY, USA, 2009. ACM.

[12] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar. Automatic scoring of online discussion posts. In *WICOW '08: Proceeding of the 2nd ACM workshop on Information credibility on the web*, pages 19–26, New York, NY, USA, 2008. ACM.

[13] J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W.-Y. Ma. Incorporating site-level knowledge to extract structured data from web forums. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 181–190, New York, NY, USA, 2009. ACM.

[14] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[15] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.

[16] L. Zhao and J. Callan. A generative retrieval model for structured documents. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1163–1172, New York, NY, USA, 2008. ACM.

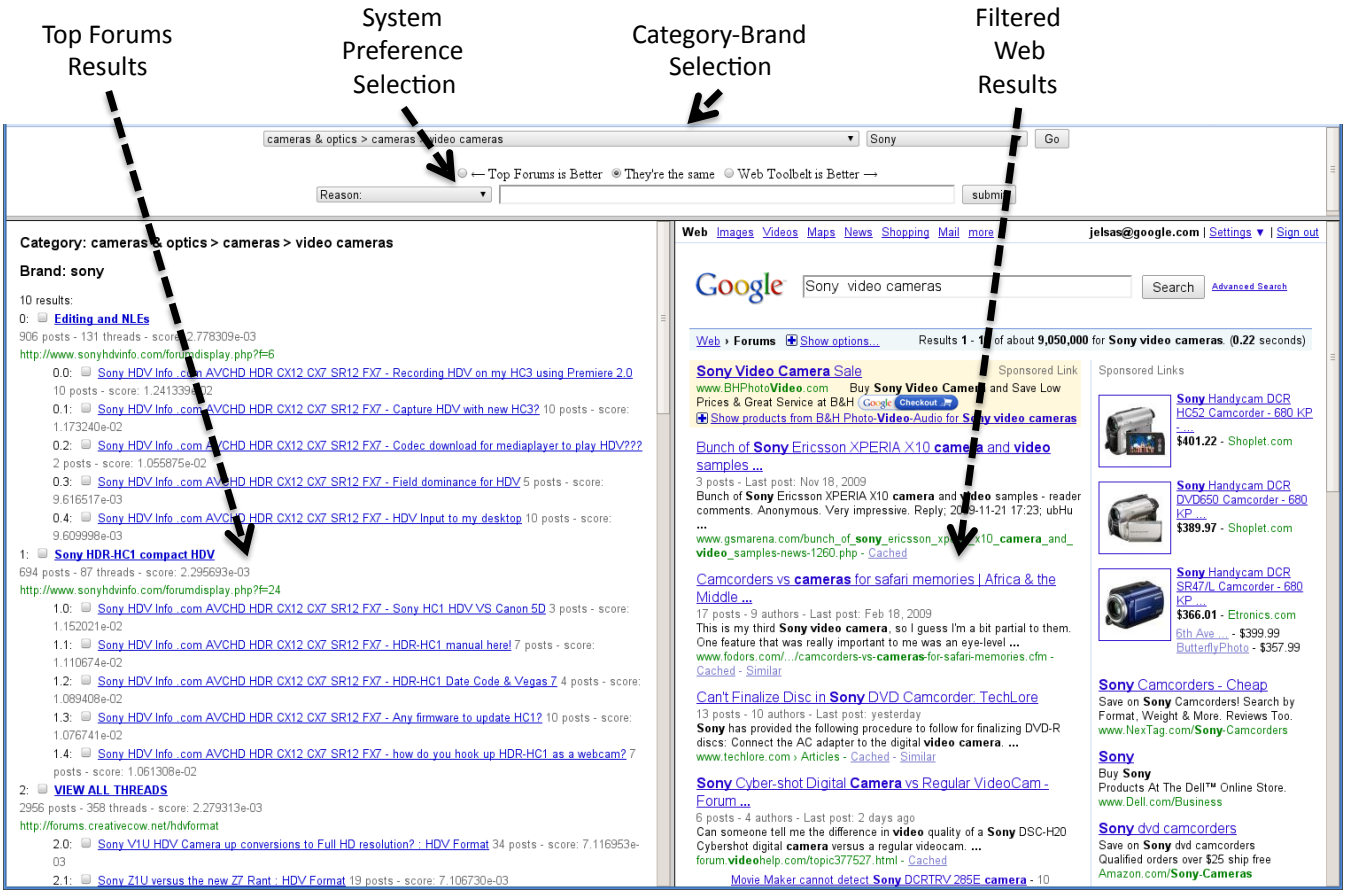

Figure 5: Evaluation interface.

Query: (Headphones, Sennheiser)

Headphones (full-size) 5963 threads - 66690 posts
www.head-fi.org/forums/f4/
- New Sennheiser models : HD465, HD485, HD201, HD215 10 posts
- Sennheiser HD580, HD580 II, HD580 Precision... what's the difference? 16 posts
- HD595 vs HD600 vs HD650 16 posts

Portable Headphones, Earphones and In-Ear Monitors 2402 threads - 26587 posts
www.head-fi.org/forums/f103/
- CX300 vs CX400 vs CX500? Or something else? 16 posts
- New Sennheiser canalphones (CX400, CX500, CXL400, CX55, CX95) 16 posts
- Sennheiser IE7: The Review (Lots of pics & comparison with IE8 & HD600 16 posts

Headphones For Sale / Trade 1564 threads - 9734 posts
www.head-fi.org/forums/f10/
- SOLD: Sennheiser HD580 w/ HD600 Grills & HD650 Cable (Australia) 9 posts
- SOLD: Sennheiser HD25-1 II w/ HD580 Cable 4 posts
- WTB: Sennheiser MX400/MX500 4 posts

Query: (GPS, Garmin)

Garmin Nuvi Forum 769 threads - 8930 posts
forums.gpsreview.net/viewforum.php?f=2
- 1390T Traffic Signal Performance vs 765T vs 265WT 2 posts
- 205W vs. 255W vs. 260W - Garmin Nuvi Forum 5 posts

www.poi-factory.com/ 3281 threads - 48519 posts
www.poi-factory.com/
- Comparing Garmin 885t with 765t | POI Factory 30 posts
- Garmin Nuvi 255W vs 260W | POI Factory 11 posts

Garmin nüvi forums 927 threads - 11538 posts
www.gpspassion.com/forumsen/forum.asp?FORUM_ID=172
- GpsPasSion Forums - Service firmware .rgn for nuvi 205w/255w/265w 7 posts
- GpsPasSion Forums - NUVI 350/360 VS ZUMO 550(Motorcycle) 17 posts

GPS Recommendations 299 threads - 2655 posts
forums.gpsreview.net/viewforum.php?f=30
- Garmin Colorado 400t, Oregon 400t, or 60Csx - GPS Recommendations 5 posts
- 255w, 255wt, 265wt??????? - GPS Recommendations 3 posts

Figure 6: Example Top Forums results from two category-brand queries, (Headphones, Sennheiser) at left and (GPS, Garmin) at right.