Deriving Wishlists from Blogs

Show us your Blog, and We'll Tell you What Books to Buy

Gilad Mishne and Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad,mdr@science.uva.nl

ABSTRACT

We use a combination of text analysis and external knowledge sources to estimate the commercial taste of bloggers from their text; our methods are evaluated using product wishlists found in the blogs. Initial results are promising, showing that valuable insights can be mined from blogs, not just at the aggregate but also at the individual blog level.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software

General Terms

Human Factors, Experimentation, Languages

Keywords

Blogs, wishlists, Amazon

1. INTRODUCTION

Blogs are a rich source of information for commercial purposes. At the aggregate level, various uses have been made of blogs and their contents, e.g., predicting spikes in consumer purchase decisions using the mere volume of blog postings [1]. We are interested in a different commercial aspect of blogs—not at the aggregate level—but at the level of individual blogs, aimed at advertisers: what offers should be made to a blogger, what advertisements should be shown to her readers, and so on.

Personal blogs provide reports about experiences and interests of individuals, objects they surround themselves with, and activities they engage in. Our working hypothesis is that from a blogger's writings we can derive her commercial taste. In this paper we zoom in on books: given a blog, our aim is to generate a list of suggestions of categories of books for the blogger to buy. We begin by mining a blog for indicators of book interests; we then use external resources to match these indicators with actual products, aggregating their category information to arrive at a book profile. We evaluate profiles thus generated by comparing them against wishlists created by the bloggers themselves.

Overall, our goal is to show that valuable insights can be mined from blogs at the individual blog level, using a combination of text analysis and powerful external resources.

2. BUILDING ADVICE MODELS

Rather than actually recommending books—recommender systems are better-equipped for this—we seek to discover the *categories* of books a blogger is most likely to be interested in, given the blog text. This enables advertisers and vendors to custom-tailor ads and offers to the blogger and her readers. Our approach to constructing this advice model is composed of two stages: first, we locate indicators for the blogger's (book) interests by mining the text; then, we use external resources to match these indicators with actual products, aggregating their category information to create a "blogger book profile."

2.1 Interest Indicators

We use two methods to identify indicators of the blogger's interests from their text: product extraction (find explicit references to books in the blog) and blog keyword extraction.

For the *product extraction* method (our baseline), we identify interest indicators by locating explicit references the blogger makes to books. Identification of names of books (and other products) is known to be difficult, and our approach to this is therefore somewhat simplistic: we tag the text with a general named entity tagger, and employ heuristics on the results to identify possible book titles. These heuristics include searching for entities in close proximity to a small set of book related keywords ("read", "book"); discarding "location" entities; matching patterns such as "⟨ENTITY⟩ by ⟨PERSON⟩," etc. Extracted entities are scored based on a combination of their recurrence in the blog, their NE-tagger confidence score, and a score derived from the heuristic used to select them; the top-scoring entities are used to populate the indicator list.[1]

For the *keyword extraction* method, we use the log-likelihood corpus-comparison method [2] to identify terms which are distinctive to a blogger: word n-grams which she uses often, compared to other bloggers. To filter out irrelevant terms extracted this way (typically, recurring proper names related to the blogger, such as family members) we ignore terms that do not appear as a noun in WordNet. All distinctive n-grams with a log-likelihood score above a threshold are taken as interest indicators.

2.2 Category Aggregation

Once a set of indicators is found, we proceed by deriving book categories matching the indicators; this is done by first retrieving categorized books that match the indicators, and then aggregating their categories into a complete profile. In practice, we use Amazon's Web Services both as a tool for locating products, and as the supplier of meta-information

[1]Cursory examination of the results of the book name identification method shows that about half of the retrieved entities are indeed books.

WWW2006, May 22–26, 2006, Edinburgh, UK.

Possible Book Titles in Text	• All My Children • Supergirl and the Legion of Super Heroes • Golden Age Superman and Earth • Roger Cline and the Peacemakers • The Bible • ...
Derived Product Profile	• Superheroes (14) • Economics (13) • Business (13) • Juvenile Fiction (11) • ...
Keywords	wolverine, replica, discretion, hulk, pencils, ...
Relevant Amazon Books	• *Wolverine: Origin* (Comics, Graphic Novels, Superheroes, Marvel, Fantasy) • *The Chaos Engine : Book 1 (X-Men: Doctor Doom)* (X-Men, Parenting & Families, Fantasy, Science Fiction) • ...
Derived Product Profile	• Superheroes (46) • Graphic Novels (39) • Fantasy (38) • Economics (37) • ...

Table 1: Product vs keyword extraction.

about them: given an indicator—a possible book name or an extracted keyword—we query Amazon for the top books related to this indicator. The categories of the returned results are aggregated and sorted by frequency in the results.

2.3 A Worked Example

We demonstrate the process of constructing the models on a particular blog: "Guided By Echoes."[2] The upper half of Table 1 shows the top-ranking book references extracted from the blog text, and the top categories associated with the generated queries to Amazon. Numbers in parenthesis are the total number of products with the listed category out of all results. Note that the extraction contains noise, for example, the (misspelled) band "Roger Clyne and the Peacemakers" which was extracted as a book. The effect of this on the final result is diminished by the fact that the Amazon queries we generate are restricted to books: in this particular example, no book results are found by Amazon.

Similarly, the lower part of Table 1 shows the keywords extracted from the blog, the top books returned by Amazon for queries containing these words, and the generated model.

3. EVALUATION

To evaluate our proposal, we require a set of bloggers for which the (book) purchase profile is known. Many bloggers point to their Amazon wishlists, which contain books and other products they desire. We downloaded a random set of 400 such blogs with their accompanying wishlist; each blog was crawled to obtain multiple posts. Non-English blogs and blogs with a small amount of text (less than 200KB, after stripping HTML and template-like text), or with fewer than 30 books in the wishlist were discarded, leaving 91 blogs with, on average, 1.1MB of text each. Wishlists were parsed in the same manner as Amazon's search results were parsed in the model construction phase: categories of books were aggregated to build a weighted list of the blogger's declared commercial interests, functioning as a golden standard. Table 2 shows this golden standard, as built for the blog used as a working example in the previous section.

Next, the methods for building advice models were employed, resulting in two models per blog: based on products

[2]URL: http://guidedbyechoes.livejournal.com.

Wishlist	amazon.com/gp/registry/17G9XYDK5GEGG
Books in wishlists	• *The Big Book of Conspiracies* (Comic books, Conspiracies, Controversial Knowledge, ...) • *Buffy the Vampire Slayer: Origin* (Young Adult, Comics, Humor, Juvenile Fiction)
Blogger Product Profile	• Games (61) • Role Playing (45) • Superheroes (42) • Comics (42) • ...

Table 2: A sample wishlist profile.

and based on keywords. To compare these models with the actual models built from the blogger's wishlists, we measured the overlap in the top-3 categories of both models: if two of the categories appearing in the top-3 model built by a method appear also in the golden model, the overlap is 2/3, and so on: in the example in Table 2 the overlap is 1/3 with both of the constructed models. In the experiments reported here we did not take into account the hierarchical structure of Amazon's categorization scheme; doing so would have resulted in higher scores—e.g., in the example, the category "Graphic Novels" is a parent category of the "golden" category "Superheroes." The average overlap over all blogs was 0.14 for the product-based extraction method, and 0.31 for the keyword-based method; experimenting with combinations of the methods did not yield additional improvements.

These initial results are encouraging: given the simplicity of our keyword method, it performs fairly well, correctly identifying about a third of the categories the blogger is most interested in, out of a large hierarchy of hundreds of different categories. An examination of failures (blogs for which no overlap exists between the models), shows that the majority of them are diary-like, highly personal blogs, with little topical substance. Often, this is computationally discernible: e.g., the keyword extraction phase for these blogs results in short lists, since only a small number of nouns exceed the minimal log-likelihood value to be considered "distinctive." A possible extension to our method would be to identify these and assign confidence values to the generated models.

4. CONCLUSIONS AND FUTURE WORK

Most computational studies of blogs focus on aggregating knowledge from a large body of text, aiming at an analysis of a given product. In this study we follow the inverse path: constructing a product profile for a given blogger. Initial results, using a simple method, suggest that this is beneficial for those targeting bloggers as consumers.

In future work, we intend to investigate "commercial contexts" in the blog—sections which are likely to be related to the blogger's desired products (such as plans for future purchases) and take the sentiment expressed by the blogger into account. Additionally, we plan to make our evaluation more robust by using the hierarchical structure of the categories, allowing for more than the exact matches we have now used. Finally, we plan to make this novel dataset publicly available to encourage additional work in this direction.

5. REFERENCES

[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05*, 2005.

[2] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *The workshop on Comparing Corpora, at ACL 2000*, 2000.