

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/4361788>

Using text mining to analyze user forums

Conference Paper · June 2008

DOI: 10.1109/ICSSSM.2008.4598504 · Source: IEEE Xplore

CITATIONS

16

READS

11,759

5 authors, including:



Ronen Feldman

Hebrew University of Jerusalem

135 PUBLICATIONS 9,019 CITATIONS

[SEE PROFILE](#)



Jacob Goldenberg

Interdisciplinary Center Herzliya

99 PUBLICATIONS 7,479 CITATIONS

[SEE PROFILE](#)



Oded Netzer

Columbia University

55 PUBLICATIONS 3,921 CITATIONS

[SEE PROFILE](#)



Lyle H. Ungar

University of Pennsylvania

498 PUBLICATIONS 25,403 CITATIONS

[SEE PROFILE](#)

Using Text Mining to Analyze User Forums

Ronen Feldman, Moshe Fresko, Jacob Goldenberg, Oded Netzer, Lyle Ungar

Abstract—Product discussion boards are a rich source of information about consumer sentiment about products, which is being increasingly exploited. Most sentiment analysis has looked at single products in isolation, but users often compare different products, stating which they like better and why. We present a set of techniques for analyzing how consumers view product markets. Specifically, we extracted relative sentiment analysis and comparisons between products, to understand what attributes users compare products on, and which products they prefer on each dimension. We illustrate these methods in an extended case study analyzing the sedan car markets.

Index Terms— Artificial Intelligence, Information Retrieval, Text Mining, Unsupervised Learning

I. INTRODUCTION

THERE is increasing recognition that product reviews by consumers can provide important insight into how they view products, and that automated text analysis methods can be fruitfully used to extract such information [1], [2]. For example, the rapidly growing field of sentiment analysis looks to extract how authors feel about different products [3]–[6]. Such work has tended to look at single products, in spite of the fact that many of the purchase decisions, and hence much of the marketing effort, is based on product comparisons. This paper describes a methodology for automatically analyzing products and comparisons between them. Given a (possibly ungrammatical) sentence such as “Sonata has soft ride similar to Camry and Accord” we automatically extract the products (Sonata, Camry and Accord) and what attributes they are compared on (“soft ride”). Our goal is to automatically determine which products are compared with each other, what attributes they are compared on, and which products are preferred on different attributes. We term this process “comparative sentiment analysis.”

The first step in comparative sentiment analysis is

extracting from discussion boards the co-occurring product names and other frequently occurring terms and the phrases they occur in. The extracted phrases provide snippets of text, similar to those in product reviews such as those provided by Zagat ®. The extracted products and attributes can be used to construct product comparison tables and graphs, which can be used either by consumers or by firms.

Even noting what pairs of products are mentioned together gives valuable insight into the structure of the market and how consumers view products. We build on previous work on entity recognition dealing with entity extraction from text and visualization of the co-occurrence of these entities in the same sentence or paragraph [7]–[9]. Entity recognition has been often used to find pairs of companies or genes that are mentioned together, sometimes in the context of a word such a “merger.” These relationships are then displayed as graphs [10]. Such work has been directed at corpora such as newswire articles or scientific abstracts. We extend these methods to the more informal styles common in consumer product reviews and blogs, and extract the type and sentiment of the relationship as well as simply co-occurrence. We locate snippets such as “...*Sonata* pricing is pretty similar to the *Fusion*”, identify the brands (shown here in italics), and extract the relations between them. We show how multi-dimensional scaling (MDS) can be fruitfully used to provide a “perceptual map” of an industry, with products being related to each other either by co-occurrence (which products are being compared) or by the similarity of the terms used to describe these products.

This paper describes an implemented system that extracts product mentions and comparisons from user discussion forums. The following sections describe the underlying methodology, and then show results obtained from a set of 868,174 reviews of cars, extracted from the sedans section of the forum at Edmund's Car Space (<http://townhall-talk.edmunds.com/WebX/ee9e22f/>).

II. METHODOLOGY FOR ANALYZING MESSAGES

Extracting structured product comparison data from product forum messages requires several processing steps. Html-pages are downloaded from a given forum site and html-like tags and non-textual information such as images, commercials, etc. are cleaned from the downloaded files. Products and product attributes are then extracted from the messages. The extracted products and attributes then allow us to visualize which products are compared against each other, showing the overall market structure, which terms are used in the comparisons, and the direction of the comparison, giving detailed insight

Manuscript received on February 22, 2008.

R. Feldman, was with the School of Business Administration in The Hebrew University of Jerusalem, Jerusalem, Israel. (phone: +972-2-5883084; fax: +972-2-5883084; e-mail: ronen.feldman@huji.ac.il).

M. Fresko, was with the School of Business Administration in The Hebrew University of Jerusalem, Jerusalem, Israel. (e-mail: freskom@huji.ac.il).

J. Goldenberg, was with the School of Business Administration in The Hebrew University of Jerusalem, Jerusalem, Israel. (e-mail: msgolden@huji.ac.il).

O. Netzer, was with Columbia Business School, Columbia University (e-mail: on210@columbia.edu).

L. Ungar, was with Computer and Information Science Department, University of Pennsylvania (e-mail: ungar@cis.upenn.edu).

into consumer preferences.

The key steps involve tagging the product mentions (information extraction), extracting snippets containing pairs of product mentions, and extracting the terms used in product descriptions of comparisons.

A. Information Extraction

We first extract the brand names (28 car companies) and model names (180 car models) from the discussion board messages. Brand names were relatively easy to extract – one can just string-match on the names – while model names showed significant variation and ambiguity. Extraction by string matching on names was not suitable for extracting models because of the variations in the writing styles. Instead, a set of regular expressions was developed for recognizing models from the model-names appearing in these informal texts [11].

The last decade has seen a profusion of papers on information extraction [7],[9], mostly using machine learning methods such as HMMs, CRFs, and other maximum entropy models, and primarily looking at formally written texts such as newswire articles or scientific abstracts [12] or partially structured web pages (e.g., seminar announcements). Entity extraction and resolution from user reviews is much trickier than from the corpora used in classic IE research. Product names vary more than personal names, and clues such as capitalization that exist in formal text are not reliable in user reviews.

Although hand-crafting extraction rules is out of fashion, we believe that there are cases, such as these informal product reviews, when it is substantially faster to build rules by hand than to label data, write feature extractors, and train models. This is particularly true since we want not only to tag terms as products, but also to resolve which product is being referred to.

The key to constructing such entity extraction tools is to note that lists of product brands and models are easy to come by on the web; the trick is to generalize the formal product names to the many variants used. Terms referring to products generally consist of a subset of the brand name, model name, and model number. Any combination of the three components can be kept or dropped. Different delimiters (e.g. space, hyphen, or nothing) can also be used between them. Brand names are mostly clean string matches to a list, while model names are often abbreviated (or misspelled). Numbers can be Roman or Arabic.

Matching is handled in two stages. First there is a tagging stage that is based on regular expressions and a list of search strings. This stage is greedy and tries to match as many terms as possible, giving high recall. After that, a second stage removes or modifies the falsely extracted entities in order to improve the precision. Based on a random sample of 500 messages and manual evaluation of the results we achieved recall of 89.4% and precision of 96.7% leading to F1 of 92.9%.

B. Snippet Extraction

We are interested in seeing which products are compared to one another, which terms show up most often in connection with a product or pair of products and what attributes consumers are using when they compare products. A term recognizer was built by using a CRF [9] model trained on the CoNLL-2000 shared task corpus [13]. It uses two consecutive CRF models - one for part of speech tagging and another for chunking. After chunking, there is filtering step, which retains only noun phrases that contain actual nouns (and not, pronouns, for instance).

The terms extracted for cars fall into several categories such as: Driving Experience (safety, aerodynamic, suspension, control, maneuverable, noise, performance, power, reliability, speed), Short and Long-Term Costs (aging, trade_in, holds_value, insurance_cost, fuel_economy, maintenance_expense, price), Internal Construction (brakes, construction, engine, platform, materials), Internal Comfort (seats, leg_room, interior_space, higher, size), and External Design (interior, design, sportive, weight, look, paint, features).

The snippet extraction component takes as input a large set of sentences in which the relevant product models are labeled. The output is a set of snippets – small sentence fragments, each containing a description of opinion, either *factual* (e.g., “Model A’s X is Y”, or “Model A has X), *sentiment-relating* (e.g. “Model A is good”, or “I like model A’s X) or *comparative* (e.g., “Model A is better than model B”, or more generally, “Model A’s X is better than model B”). We focus here on the comparative snippets.

Snippet extraction is done using the following steps:

Preprocessing:

The texts are tagged with parts of speech (PoS) and chunked into noun phrases using a CRF-based tagger and chunker, both trained on CoNLL-2000 shared task training set. During preprocessing, we also find lists of models (“Models X, Y, and Z”) and convert them into a single term, so that the system can more easily extract opinions expressed about multiple models simultaneously.

Pattern Extraction:

We generate the set of surface patterns by a suitable modification of the A-priori association mining algorithm. Each such pattern is a sequence of tokens including the special slot-mark tokens for product names and (optionally) skips, which indicate gaps in the pattern. First, we extract all sequences (without gaps) of tokens, PoS tags, and noun phrase (NP) chunks that appear in the set of all sentences with frequency greater than a given minimal support value. Then we mine the sentences (as ordered sets of such sequences) for frequent item-sets. The result is the set of all sufficiently surface patterns.

Pattern Filtering:

In order to reduce the number of irrelevant patterns, we

keep only patterns that contain specified words. In the first set of results, for the two-model case, we require at least one of the words "more", "less", or "than", or at least one adjective. Also, all of the patterns must start with a model. If a pattern contains two models, then the pattern must also end with a model. We also tried simple parsing, but we found that fairly simple pattern matching works better than an off-the-shelf parser. Empirically, sentences are often ungrammatical, and comparisons in these corpora often span sentence boundaries, but generally fit the pattern *model ... comparative than ... model*.

Snippet Extraction:

Snippets from the user reviews are extracted by matching the patterns against the corpus. Matching is conceptually straightforward. We use a fast algorithm that finds all matches of all patterns with complexity logarithmic in the number of patterns and linear in the corpus size. We used two very general patterns to extract snippets:

<Model> * /JJ than * <Model>

<Model> * /JJ * to * <Model>

where /JJ means "adjective" and "*" means a gap is allowed. This produced 3,025 snippets. When the snippets are extracted, we obtain the pairs of products and the comparative term used between them. This information can be extracted and put into a table or a graph. We describe such analysis below.

III. EXAMPLE RESULTS

Starting with 868,174 messages containing 5,972,695 sentences, we extracted 28 car brands, 180 car models and 1,037 different terms. There were 20,768 mentions of brand, 575,110 of car models, and 6,194,507 of terms.

A. Co-occurrence Results

One can examine the comparisons consumers make at the brand or model level to draw a "perceptual map" of the similarities between car companies. We looked at several different ways of displaying co-occurrence data, and found that the most informative is plotting brands using multi-dimensional scaling (MDS) [20] on co-occurrence lift -- the ratio of the actual appearance of two nodes together to the frequency we would expect to see them together. Using MDS on the lift of the comparisons between each pair of the 28 car brands mentioned in the sedan cars forum we create a perceptual map of the car brands, where the axes represent perceptual dimensions inferred solely from co-occurrence of mentions in user comments. As can be seen in Fig.1, as we go from left to right the brands seem to be decreasing in prestige, and from bottom to top the brands decreasing in the age of the target market they appeal to. Furthermore, similarly perceived cars are clustered together in this space. On the left of the map we find the luxury car brands: the Japanese: Acura, Lexus and Infiniti on the top, the German: BMW, Audi and Mercedes in the top middle, the other European: Volvo, Saab and Jaguar in the bottom middle, and the two American brand that seem to be perceived as more luxurious (Lincoln and Cadillac) at the

bottom. On the right of the figure, from top to bottom, we find the high share Japanese cars (Honda, Mazda, Nissan and Toyota) followed by some Korean and lower market share Japanese in the middle, to the majority of the American brands at the bottom. Note that Ford, which has a stronger international presence, is positioned closer to the non-US brands. The lower share Japanese Mitsubishi and Suzuki brands are clustered together with the Korean cars.

Such a perceptual map could be very informative in determining the market structure and competitive landscape as well as in understanding the word-of-mouth pattern expressed in the forum. The perceptual map seems to provide strong face validity to the information provided from brand comparisons in forums. While for cars, this analysis provides few surprises, for less well-studied product categories, such an analysis might be far more informative.

We also built an MDS map using the raw co-occurrence data, and using the terms mentioned in connection with each car. The maps generated in each case share gross characteristics but are different in their details.

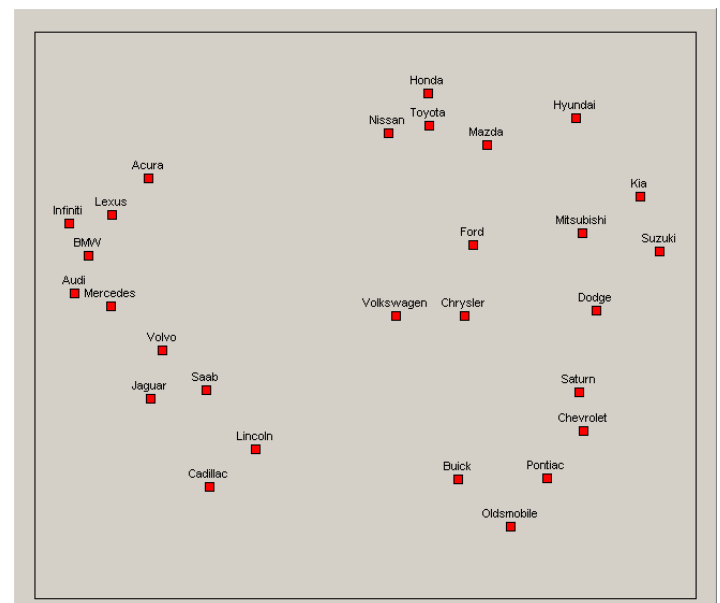


Fig. 1. A Perceptual Map of Car Brands

Snippets were classified into three categories, depending on whether the pair of cars was judged to be (a) similar, (b) different or (c) neither. Similarity of products is fairly reliably indicated by the word “similar” (e.g., “*Sonata* has soft ride similar to *Camry* and *Accord*”), while differences between products rarely mention the word ‘different’; they say how it is different (e.g., “*300 C Touring* looks so much better than the *Magnum*”, “I drove the *Honda Civic*, it does not handle better than the *TSX*, not even close”). Many snippets neither indicate similarity nor difference, but merely mention a pair of products, for example asking how they compare.

We tried training a support vector machine to predict the category labels of the snippets. We manually labeled 200 snippets each containing pairs of car models and of shoe models. (We found a similar set of user comments on shoes in *Runner's World Magazine Discussions* in <http://forums.runnersworld.com/eve/forums/a/frm/f/68710647/7/p> and used these to see how well models developed for one product on one web site generalize to a different product on a different site.; See [11] for a description and more analysis of the shoe discussions.) Training on cars and testing on shoes gave a micro-averaged F1 of 0.66; training on shoes and testing on cars gave an F1 of 0.76. (Coincidentally, precision was equal to recall in both cases, so these are also the precision and recall accuracies.) We also wrote a very simple rule-based classifier of the following form: label as similar if it contains the words “similar” or “comparable” or as different if it contains the words “differ”, “more”, “less” or any comparative adjective. All other snippets are labeled as “other”. These gave significantly better micro-averaged F1's of 0.785, and 0.82 on the same shoe and car test sets.

Focusing in on the snippets classified as “different,” we can easily extract from those snippets which car or shoe was labeled as better, and the keyword or phrase (if any) indicating on what attribute they were better. The result is a count of occurrences of 4-tuples of the form:

4 compare(Camry, Accord, noise, better).

This should be interpreted as “there were 4 snippets in which the Camry was judged to be better than the Accord in terms of noise”.

A summary of the results for comparisons between the Accord and Camry gives the flavor of what is learned. The “vote” is 12 to 8 saying the Accord is better than the Camry (these are general comparisons, where no attribute was found) but there is almost an even split on which has a better price. The count is 5 to 0 saying the Accord has worse noise than the Camry and 3 to 0 saying the Accord has a better interior. The Accord is perceived to be more sportive, and better to drive, but the Camry is bigger, has more leg room and looks better.

For objective facts such car theft, size, or sales, one can ask how accurate the sentiment is. For example, we compared user sentiment about cars being stolen with the reported rates at which the cars were stolen. The **National Insurance Crime Bureau** (NICB®) has compiled a list of the 10 vehicles most frequently reported stolen in the U.S. in 2005. Six of these appear in the top 10 cars mentioned with “stealing” phrases in our data (“Stolen,” “Steal,” “Theft”). The top five from each

TABLE I
COMPARISONS FOR TOP 5 CARS

National Insurance Crime Bureau	Consumer Discussion Board
1) 1991 Honda Accord	1) Honda Accord (165)
2) 1995 Honda Civic	2) Honda Civic (101)
3) 1989 Toyota Camry	3) Toyota Camry (71)
4) 1994 Dodge Caravan	4) Nissan Maxima (69)
5) 1994 Nissan Sentra	5) Acura TL (58)

source are shown in Table I; the left column gives the NICB stolen car list; the right column gives our extracted “stealing” brands, along with the number of times they were mentioned in connection with stealing.

IV. CONCLUSIONS

As is clear from the growing scholarly [1],[3],[4],[5], [16]-[18] and commercial interest in sentiment analysis (commercial companies include BuzzMetrics, Reputica, Umbria, Cymfony, BuzzLogic, and SentiMetrix), discussion boards provide a potentially attractive source of information about how consumers view products. None of these researchers except for [19], [21] focused on the relationships among or comparisons between different products. We believe that in addition to the more widely studied work on sentiment analysis for individual products, useful insight can be gained by extracting pairs of products. At the highest level, graphs of co-mentions of pairs of products give an overview of one type of market structure: what is compared against what. MDS on co-occurrence lift provides good visualization of the market. At a more detailed level, comparative sentiment analysis provides insight into what attributes products are compared on, and which products are preferred to others.

At several points in the extraction process we compared state-of-the-art machine learning methods against simple hand-coded patterns. Although CRFs were useful for part-of-speech tagging, and the part of speech tags were useful in extracting comparisons, it was often the case (e.g., when resolving forms of mentions of model numbers to the actual model, or in classifying messages into “similar”, “different”, or “other”) that surprisingly simple rules worked well, and worked well for the disparate domains of shoes and cars. No doubt, if we used the right feature set and more training data, better results could have been obtained with CRFs or SVMs, but the point is that the right features (e.g., the presence of a comparative adjective or of an Arabic or Roman numeral) make the deciding difference in the quality of the results. Note that this contrasts the dominant direction of current research in information extraction [22]. User discussion boards differ significantly from the text used in competitions such as MUC, TREC and CoNLL [13], which focused on extracting entities (typically people, places and organizations) from relatively carefully written formal text such as that of newswire articles. Although more careful study is required, preliminary experiments suggest that parsing is far less helpful on informal

texts. As a further difficulty, there does not appear to be any good collection of resolved informal product model names that can serve as a basis for supervised learning methods. We thus rely more on hand-built regular expressions.

Marketing campaigns are often designed in part based on product comparison tables: what is my product compared against, and on what dimensions? Extracting co-mentioned pairs of products and the terms that show up frequently with them provides precisely this kind of table. Cars are a widely studied and a relative stable market, providing nice face validity to our extraction system. However, for markets such as computer games, electronics, movies and music, which change rapidly, one can utilize our method to better analyze and understand the market. We suspect that for such dynamic markets the attributes used as the basis of comparison such as disk size, latency, interface and color may become more or less important as differentiators. In such cases, even knowing what dimensions products should be compared on is nontrivial. Even in relatively low tech markets such as running shoes, it is common to see sites such as Sports-Authority offering comparisons of 65 shoe models, with no mechanism for selecting a subset other than picking a brand. With evidence accumulating that offering more choices can lead to lower consumer satisfaction and lower purchase rates [23], it is becoming ever clearer that picking the right set of products for comparison is of great value for both consumers and retailers. Mining consumer product reviews provides a relatively inexpensive mechanism not only for selecting potential products to compare against a given product, but also the dimensions on which to compare the products on.

ACKNOWLEDGMENT

We thank Benjamin Rosenfeld for helpful comments on earlier drafts of this paper.

REFERENCES

- [1] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," *Proceedings of HLT-EMNLP*, 2005, pp. 339–346.
- [2] K. Dave, S. Lawrence and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proceedings of the 12th Int. World Wide Web Conference WWW-2003*.
- [3] S.-M. Kim and E. Hovy, "Identifying and Analyzing Judgment Opinions," *Proceedings of HLT/NAACL-2006*, New York City, NY, pp. 200–207.
- [4] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of ACL-2002*.
- [5] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.
- [6] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Morristown, US, 2002, pp. 79–86.
- [7] D. Miller, R. Schwartz, R. Weischedel and R. Stone, "Named Entity Extraction from Broadcast News," *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, 1999.
- [8] R. Feldman, B. Rosenfeld and M. Fresko, "TEG – A Hybrid Approach to Information Extraction," *Knowledge and Information Systems*, vol. 9, issue. 1, Jan. 2006, Springer-Verlag, New York, USA., pp. 1–18.
- [9] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 188–191.
- [10] R. Feldman, W. Kloesgen and A. Zilberstein, "Visualization Techniques to Explore Data Mining Results for Document Collections," *Proceedings of KDD-97*, 1997.
- [11] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer and L. Ungar, "Extracting Product Comparisons from Discussion Boards," *Proceedings of ICDM-2007*, Omaha, NE, 2007.
- [12] A. Yeh, L. Hirschman, A. Morgan and M. Colosimo, "BioCreative task 1a: Gene Mention Finding Evaluation," *BMC Bioinformatics*, 2005, 6.
- [13] E. F. T. K. Sang and S. Buchholz, "Introduction to the CoNLL-2000 Shared Task: Chunking," *Proceedings of 4th Conference on Computational Natural Language Learning*, 2000, pp. 127–132.
- [14] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld and A. Yates, "Unsupervised Named-Entity Extraction from the Web: an Experimental Study," *Artificial Intelligence*, 2005, 165(1), pp. 91–134.
- [15] B. Rosenfeld and R. Feldman, "Using Corpus Statistics on Entities to Improve Semi-supervised Relation Extraction from the Web," *Proceeding of ACL-07*, 2007.
- [16] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," *Proceedings of CIKM-05*, Bremen, DE, 2005.
- [17] T. Chklovski, "Deriving Quantitative Overviews of Free Text Assessments on the Web," *Proceedings of 2006 International Conference on Intelligent User Interfaces (IUI06)*, Sydney, Australia, 2006.
- [18] M. Gregory, N. Chinchor, P. Whitney, R. Carter, A. Turner and E. Hetzler, "User-Directed Sentiment Analysis: Visualizing the Affective Content of Documents. In: Sentiment and Subjectivity in Text," *Workshop at the Annual Meeting of the Association of Computational Linguistics ACL-2006*.
- [19] N. Jindal and B. Liu, "Mining comparative sentences and relations," *Proceedings of AAAI'06*, 2006.
- [20] J. M. Lattin, P. E. Green and J. D. Carroll, "Analyzing Multivariate Data," Thomson Brooks/Cole, Pacific Grove, CA., 2003.
- [21] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," *Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 244–251, 2006.
- [22] A. McCallum, "Information Extraction: Distilling Structured Data from Unstructured Text," *ACM Queue*, vol. 3, number. 9, Nov. 2005.
- [23] B. Schwartz, "The Paradox of Choice: Why More Is Less," HarperCollins Publishers Inc., New York, NY10022, 1st ed., 2004.