Towards Automated Reputation and Brand Monitoring on the Web

Cai-Nicolas Ziegler Michal Skubacz

Siemens AG, Corporate Technology IC 1
Otto-Hahn-Ring 6, D-81730 München
{cai.ziegler, michal.skubacz}@siemens.com

Abstract

The ever-increasing growth of the Web as primary provider of news and opinion shaper makes it impossible for individuals to manually spot and analyze all information of particular importance for globally-acting large-scale corporations. Hence, automated means of analysis, identifying upcoming topics of specific relevance and monitoring the reputation of the brand at hand, as well as its competitors, are becoming indispensable.

In this paper, we present our platform for analyzing Web data for such purposes, adopting different semantic perspectives and providing the market analyst with a flexible suite of instruments. We focus on two of these tools and outline their particular utility for research and exploration.

1 Introduction

The advent of consumer-generated media, such as blogs and newsgroups, and the proliferation of online news channels on the Web make traditional market monitoring virtually impossible. Hence, there is an emerging trend towards automated market intelligence [4] and the crafting of tools that allow reputation monitoring in a mechanized fashion.

We present one such approach that is based on the massive collection of various document types from the Web, e.g., HTML pages, RSS feeds, and newsgroup messages, all gathered on a regular basis. Our market intelligence platform addresses the information needs of large corporations, analyzing news sources through text mining techniques and providing information views that allow for the detection of upcoming topics and the observation of likely competitors.

To this end, we briefly describe our data collection and storage architecture and then move on to describe two selected analyses that are integral parts of our platform. The first one allows the visual exploration of relevant documents through the creation of keyword networks. Owing to the transitivity of links, discovering new topics and key players in the network is facilitated. The second analysis allows to directly match two competitors against each other, with respect to their Google rank and the DMOZ Open Directory categories their respective Web representations fall into.

2 Related Work

Probably the most prominent representative for approaches harnessing text and Web mining techniques for market intelligence is IBM's WEBFOUNTAIN project [6], based on massive server clusters and providing information to service clients in an ASP-like fashion. IntelliSeek's BRANDPULSE [4] and BLOGPULSE [5] platforms pursue similar objectives, but focus primarily on consumer-generated media, such as weblogs, rather than arbitrary Web content. Morinaga *et al.* [7] present an approach that automatically mines consumer opinions with respect to given products, in order to facilitate customer relationship management. Agrawal *et al.* [1] and Gamon *et al.* [3] have also conducted research in opinion mining from Web corpora for marketing purposes. To this end, NLP techniques have also been tried in order to complement text mining efforts [9].

3 Information Collection and Indexing

The section at hand gives a brief overview of our information collection and storage architecture, which provides the foundations of all sophisticated mining and advanced text analysis processors running on top. Data collection is subdivided into two primary tasks, namely gathering documents bearing content, e.g., news, and collecting background knowledge, e.g., classification taxonomies, that serve as background knowledge for all our analyses.

3.1 Content

Low-level operations are assumed by link extractors that process various information sources on the Web on a regular

and/or on-demand fashion, extracting links for later crawling and analysis. For instance, we have RSS feed extractors that query approximately $340,000$ RSS feeds every day, extracting links to feed messages that have not been collected so far. The seed RSS feed lists are provided by continuously monitoring RSS and blogging directories, such as SYNDIC8 (*http://www.syndic8.com*), 2RSS ((*http://www.2rss.com*), or BLOGSTREET (*http://www.blogstreet.com*). These directories also feature taxonomic information the RSS feeds are categorized into.

Several extractors forward pre-defined searches to popular search engines and record the top-N results returned, in order to obtain more focused and ranked results. HTML pages and newsgroups are also harvested through spiders.

Next, multi-threaded crawlers fetch entries from the queue of links not processed so far and download the respective HTML, XML, or NNTP content. Basic processing is applied in order to strip tags, comments, script blocks, commercials, and so forth. The results are then inserted into an inverted index for fast access. High-level analysis tasks operate both on the raw documents as well as the index.

A warehouse-like data cube is constructed in the last step, organizing the collected message facts according to diverse criteria, e.g., date of publication, content type description, categorization, and so forth. For each dimension, various hierarchy levels are provided, enabling the user to drill down or roll up when analyzing new media coverage with our OLAP-style data visualization platform[1]. The cube is constructed in an ad-hoc fashion, based on monitoring tasks the user can specify.

3.2 Metadata

Besides news, weblogs, and other Web documents, our information architecture also comprises various databases with massive background knowledge. For better categorization, we have collected diverse RSS feed classification hierarchies from RSS directories, requiring that all our feeds are arranged into at least one leaf node of those. Merging of classification taxonomies has been done manually.

For search result classification, in particular results from Google, we have collected the taxonomy from the Open Directory Project (ODP). This directory, made up of 590,000 categories at the time of this writing, represents the joint effort of thousands of voluntary editors.

For entity recognition, we use several tables containing person names, sorted according to their frequency and annotated with the country of provenance, companies, organizations, and geographic locations. For the geographic locations alone, we have more than 7 million entries. Extensions are envisioned that hold particular domain knowledge, e.g., names of CEOs, annotated with the branch of industry, or the names of major player corporations for selected branches.

[1] Owing to space constraints, the platform is not described in this paper.

1	John Reed	11	Goldman Sachs
2	Richard Grasso	12	AIG
3	Jack Welch	13	Traders
4	David Weidner	14	Home Depot
5	Harvey Pitt	15	Interim
6	Dennis Kozlowski	16	Citigroup
7	Head	17	New York
8	Bell	18	Stock Exchange
9	Enron	19	Chairman
10	Investor	20	NYSE

Figure 1. Top-20 results for DICK GRASSO

4 Navigating Networks and Finding Needles

The first analysis and result visualization instrument we describe is an exploratory tool for extracting navigable information networks from sets of documents. The tool is geared towards providing different semantic views of the data, helping the user to discover new interrelations (e.g., of people, corporations, etc.) through maintaining the network perspective.

4.1 Tool Outline

For each analysis task, a center query and context query need to be defined. The center query phrase serves as the network's origin, while the context query only serves as means to filter non-relevant document, which may prove useful when the center query is ambiguous. All documents satisfying both queries are selected and each n-gram within a defined term distance from the center query is inserted into a candidate list. Candidate phrases satisfy a parameterizable support threshold, which can be used for efficient n-gram computation, exploiting the apriori property [2].

Parameterizable pre-processing filters can be plugged into the candidate list generation process. Unwanted terms and phrases can thus be removed from the outset. Typical examples include stopword filters, POS-taggers for discarding certain parts of speech, and so forth.

Upon creating the list with candidate n-grams, post-processing is applied through various weighting filters than can be plugged into the system dynamically. These plug-ins may penalize or reward candidate n-grams; the weights that each plugin assigns are normalized via z-scores across all candidate n-grams. For instance, we have weighting plugins for person names, corporation and organization names, frequency of occurrence counting, and also for phraseness and informativeness of terms [8].

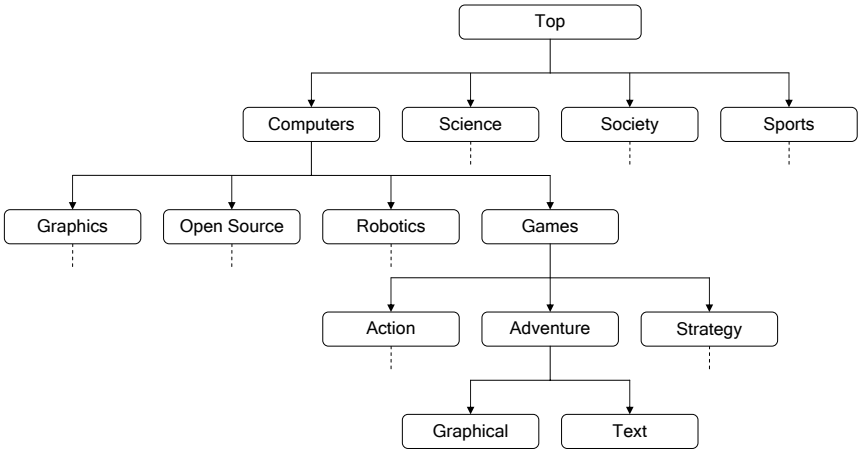

Figure 2. Extracted fragment from the DMOZ taxonomy

The eventual score of each candidate term is the sum of all weights assigned by the post-processing filters. Hereby, each filter's impact can be increased or decreased. Finally, the top-M n-grams for the center phrase are selected as winners. The results for sample query DICK GRASSO, ex-NYSE chief, are displayed in Figure 1. In this case, we have boosted the person name filter with factor 50, and company names with factor 25. Some terms not being named entities still made it into the list owing to their high frequency of co-occurrence, e.g., HEAD.

When visualizing the network, the top-M n-grams become the direct neighbors of the node representing the center phrase. The whole procedure is then repeated recursively for these neighbors, becoming center phrases themselves.

4.2 Practical Use

The flexible plugin mechanism makes our exploratory analysis a versatile instrument that can be used for a broad range of monitoring tasks.

Extracting social networks is one key application scenario, e.g., for knowing which persons the managers involved with a merger interact with, possibly across multiple transitive links of co-citation. Hereby, our analysis application allows the user to select the edges connecting persons, so as to obtain all messages and blogs where these two were mentioned together.

Another interesting use is that of brand awareness, i.e., analyzing which descriptive words are used in the context of a given brand or product. To this end, plugins for filtering terms by their phraseness and informativeness [8] are applied.

5 Taxonomies for Matching Brand Profiles

The second reputation monitoring tool we present serves as means for matching the Web perceptions of two competing corporations, e.g., NOKIA and BENQ, against each other. To this end, we leverage the power of Google Directory (*http://www.google.com/dirhp*), an additional service of the popular Google search engine, and the DMOZ Open Directory Project (*http://www.dmoz.org*). At the time of writing, DMOZ is probably the most comprehensive human-crafted taxonomy so far, being made up of 590,000 hierarchically arranged categories that classify more than 4 million Web sites.

5.1 Approach Implementation

Google Directory delivers search results for a given query in the usual fashion, but attributes each result with the DMOZ taxonomy node the latter falls into, e.g., CONSUMER INFORMATION → ELECTRONICS → ... → MOBILE. For each of the two queries, i.e., NOKIA and BENQ in the case depicted, we collect the top-N search results and, more importantly, their corresponding referrals into the DMOZ taxonomy. These N referrals are then used to build a semantic fingerprint of each query, according to the probabilistic score propagation scheme proposed in [11], [10] and [12]. The closer one given search result to the top rank, the higher its impact on the semantic profile[2]. The latter feature of our approach is based on the hypothesis that top results are more characteristic for a query than results further down the list, which may represent only side aspects of minor importance.

[2]We use exponential half-life for decreasing ranks.

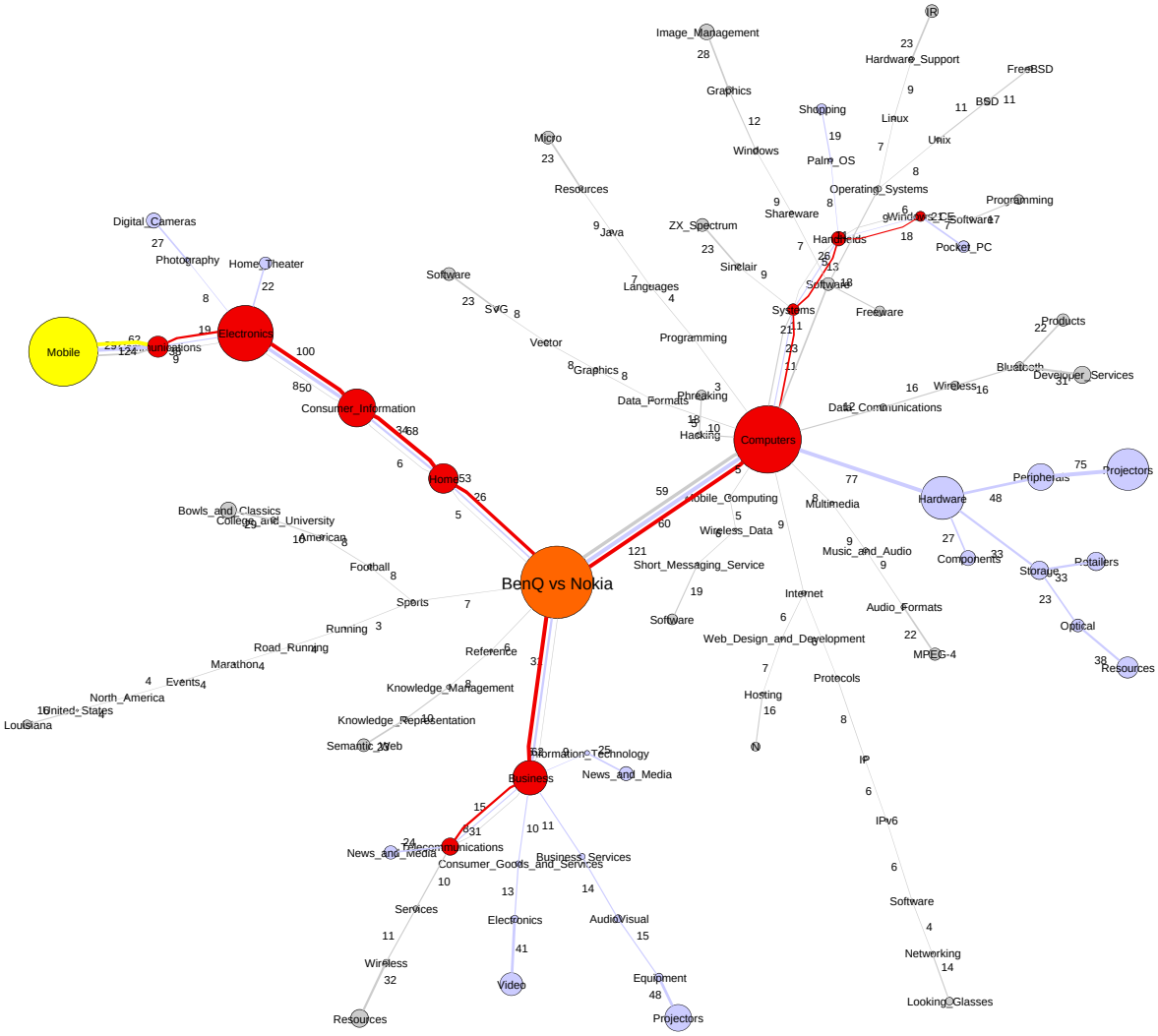

Figure 3. Semantic profile matching for BENQ and Nokia

The semantic profiles for each query, such as the name of a corporation or brand, represent the categories the associated concepts are classified into. Since higher-ranked results have an increased impact on the profile's makeup, corporations are described according to their principal lines of business.

Next, we transform the semantic fingerprint of each of the two queries into a score vector and compute Pearson's correlation coefficient in order to determine the semantic similarity between them. As a result, we obtain a score taken from the interval $[-1, +1]$, where greater values denote higher similarity.

5.2 Application Scenarios

However, besides the mere scalar similarity score, we also obtain extensive information about the taxonomy locations where the overlaps of both profiles occur. Moreover, these locations are assigned weights, denoting the strength of both concepts' communality. Figure 3 depicts the DMOZ categories where overlaps between NOKIA and BENQ, two corporations primarily dealing with telecommunications. As can be seen, the strongest communality is in mobile consumer electronics, but also in computer hardware components, Pocket PCs, and digital cameras.

The analysis tool can be used for various monitoring scenarios. First and foremost, it enables the market analyst to

verify the compliance of a brand's Web perceptions with its real business focus, in particular with respect to major competitors. Deviations from desired behavior can be tracked easily, enabling immediate action-taking.

6 Outlook and Conclusion

In this paper, we have outlined the information infrastructure foundations of our reputation monitoring platform, giving more detailed background information for two of our analysis tools and their use.

At the time of this writing, we are working on sentiment detection modules [9, 7] for complementing the platform. Our declared goal is to create a lightweight tool suite that can be tailored to the specific monitoring problem at hand, rather than offering one huge text processing facility as is the case for IBM's WEBFOUNTAIN.

References

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference*, pages 529–535, Budapest, Hungary, 2003.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann.

[3] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Proceedings of the 6th International Symposium on Intelligent Data Analysis*, pages 121–132, September 2005.

[4] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 419–428, Chicago, IL, USA, 2005.

[5] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *Proceedings of the WWW 2004 Workshop on the Weblogging Ecosystem*, New York, NY, USA, 2004.

[6] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal*, 43(1):64–77, 2004.

[7] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the Web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 341–349, Edmonton, AL, Canada, 2002. ACM Press.

[8] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Workshop on Multiword Expressions of ACL*, Sapporo, Japan, July 2003.

[9] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 427–434, Washington, DC, USA, 2003. IEEE Computer Society.

[10] C.-N. Ziegler, G. Lausen, and L. Schmidt-Thieme. Taxonomy-driven computation of product recommendations. In *Proceedings of the 2004 ACM CIKM Conference on Information and Knowledge Management*, pages 406–415, Washington, D.C., USA, November 2004. ACM Press.

[11] C.-N. Ziegler, S. McNee, J. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, May 2005. ACM Press.

[12] C.-N. Ziegler, K. Simon, and G. Lausen. Automatic computation of semantic proximity using taxonomic knowledge. In *Proceedings of the 2006 ACM CIKM Conference on Information and Knowledge Management*, Washington, D.C., USA, November 2006. ACM Press. to appear.