Finding your way through blogspace: Using semantics for cross-domain blog analysis

Bettina Berendt and Roberto Navigli
Humboldt University, Berlin, Germany, and Università "La Sapienza", Roma, Italy
http://www.wiwi.hu-berlin.de/~berendt, and http://www.dsi.uniroma1.it/~navigli

Abstract

Blogspace is one of the most dynamic areas of today's Internet, and it is increasingly recognised that blogs are much more than "meaningless chatter". Many syntax-based approaches exist to analyse the text and the network structure between blogs. While this is very helpful for purposes such as the detection of discussion bursts concerning uniquely-named topics (e.g., a book, product, or person), it is insufficient for understanding blogs discussing new phenomena in different wordings, or for finding and explaining relationships between new discourse topics or the context of a new topic in a larger domain of discourse. In this paper, we propose two methods for semantics-enhanced blogs analysis that allow the analyst to integrate domain-specific as well as general background knowledge. The methods rely on the Term Extractor for identifying keyphrases (Navigli & Velardi, 2004), SSI (Structural Semantic Interconnections) for disambiguating terms (Navigli & Velardi, 2005), and the taxonomy of domain labels by (Magnini & Cavaglià, 2000). Applications include topic detection and grouping, the proposal of blog tags and the forming of blog directories, and blog recommender systems. To illustrate the usefulness of our approach, we present a detailed experimental analysis of a sample of four sets of blogs with different thematic foci (food, health, law, and weblogs about blogging).

Introduction

Blogspace is one of the most dynamic areas of today's Internet and global communications network, and it is increasingly recognised that statements published in blogs are much more than "chatter". Mainstream media have elaborated on issues that were initially raised in blogspace[1], and journalism prizes have been awarded to outstanding blogs[2]. Marketers are hoping to tap blog-published consumer opinions on brands and products as a leading indicator of purchasing behaviour and customer satisfaction. Data mining appears to be the most promising methodology that can leverage the electronic publication medium to understand phenomena in blogspace and to use them for prediction.

Most blogs are characterised by a combination of two features that arguably make blogspace a new medium: a focus on publication (as opposed to discussion as in newsgroups), and a strong sense of community (as opposed to one-way communication in broadcast publications). Much research has focused on the community aspect and investigated the network characteristics of blogspace and their dynamics. The publication aspect has been recognised as an important indicator of new topics of discourse, leading to the utilisation of topic detection and tracking methods in blog analysis.

However, the approaches so far have relied purely on syntactic-statistical methods. While this is very helpful for purposes such as the detection of an upsurge in discussion on a uniquely-named topic (such as a new book or other product in the marketing domain, or a person in the social-political domain), it is insufficient (a) for detecting the emergence of new phenomena that are discussed with different wordings within or across communities, (b) for relating new discourse topics to each other, or (c) for placing new discourse topics into the context of larger domains of discourse.

In this paper, we propose two methods for semantics-enhanced blogs analysis that allow the analyst to integrate domain-specific as well as general background knowledge. Applications include semantic topic detection [(a) above] and semantic topic grouping (c), both of which can help to propose tags to blog authors and form blog directories, and help blog readers by more powerful blog searches. A further application is blog recommendation that can react immediately to an onset of a relevant publication activity in a blog that was hitherto unknown to a reader (b).

The methods rely on the utilisation of our Term Extractor for identifying relevant keyphrases (Navigli & Velardi, 2004), SSI (Structural Semantic Interconnections) for disambiguating terms (Navigli & Velardi, 2005), and the taxonomically-structured domain labels from IRST (Magnini & Cavaglià, 2000). The methods use general-purpose semantics (WordNet, see Fellbaum, 1998), domain-specific knowledge (the IRST domain labels[3]), and statistical measures (co-occurrences and term frequencies in se-

[1] e.g., "Rathergate", cf. www.salon.com/opinion/feature/2005/-03/09/rather/; see also blogresearch.com/ref.htm

[2] e.g., www.grimme-online-award.de/de/preistraeger/-preistraeger_2005/bildblog.htm

[3] IRST domain labels are available at: http://wndomains.itc.it

lected corpora); they demonstrate the power of combining statistical-syntactical with semantic analyses.

We illustrate the usefulness of our approach by an analysis of four thematic groups of blogs, their semantic description, and the detection of new relationships within and between the blogs and their content.

The paper is organised as follows: In the next section, we describe related work and motivate the semantics-based approach. After that, we describe the experiments, focusing first on the data, then on a baseline syntactic analysis, and then on two types of semantics-enhanced analyses. We explain the methods and tools used along with the experiments: the Term Extractor, the domain label taxonomy, SSI, and path grammars for filtering out interconnections that induce topic drift. We give a detailed quantitative and qualitative description of the results, and conclude with a summary and outlook on future work.

Related work

The analysis of blogs has focused on two main issues. One is the analysis of a set of blogs and the explicit network constituted by the hyperlinks, blogrolls, trackback links, etc. contained in its documents, and the other is the analysis of the texts in single blogs or blog corpora.

Explicit blog networks are studied in the tradition of social network / Web community analysis, using methods from Web structure mining, e.g., (Adar et al., 2005). Examples include (Adamic & Glance, 2005) who studied the interconnectedness of Republican and Democrat blogs and identified different degrees of interconnectedness of the two communities. Also relevant is the study (Borgs et al., 2004) that used the hyperlink structure of cross-posts and hierarchical text clustering to find a "hidden network" of related topics in newsgroups.

However, such blog networks are the result of many human decisions to set a hyperlink or a trackback link, or to include another blog in one's blogroll. This works well for very active and knowledgeable bloggers, and it creates densely connected, high-quality blog networks. Exaggerating somewhat, we could say that these bloggers / blogs already *are* a community, and science analyses these communities. But how can science help (especially new) bloggers to *become* a community? Answers to this question generally rely on the textual content of blogs.

Prior to supporting authors, text analysis is employed to understand blogs. One important application is topic detection and tracking. This is interesting for descriptive studies such as (Adamic & Glance, 2005) who studied major Republican and Democrat blogs in the run-up to the 2004 US presidential election, finding out, for example, what persons and topics were most talked about in both groups. High hopes are currently focused on the predictive analysis of discourse topics (Glance et al., 2005; Yi, 2005), for example in marketing where discussions of books and their authors have been shown to be a leading indicator of spikes in book sales (Gruhl et al., 2005). Text analysis focuses on syntactic-statistical methods such as bigram finding, frequency thresholding, and filtering by part-of-speech as well as by the "burstiness" within a time series of phrase occurrence (Glance et al., 2004), employed in the `www.blogpulse.com` software and Web interface.

As in the Web at large, metadata are accepted as a means to increase the quality of blog understanding, grouping, and search engine visibility. Usually, *tags* are added manually to a blog. In most systems, tags are freely chosen by the user; their technical realisation and the grouping of blog posts into the tag directory is supported by blogging software. This allows the user to help create participatory media. The well-known blog tagging site `www.technorati.com` currently (Jan. 2006) tracks close to 5 million tags, with the number increasing by about 1 million per month. These tags are not part of a controlled vocabulary, and they are not aggregated into a hierarchy. (Compare this with the current state of the `www.dmoz.org` / Google directory: more than 70,000 editors have categorised more than 5 million sites, agreeing on below 600,000 categories organised into a hierarchy with only 16 top-level categories.)

This "folksonomy" (collaborative categorisation using freely chosen keywords) stands in contrast to established classification methods in library science that rely on controlled vocabularies, taxonomical organisation, and/or given analytical facets. Folksonomy has been argued to be superior for fast-changing domains with unclear boundaries (for an introduction and comparison, see Quintarelli, 2005). However, the advantages hinge on large numbers, of people as well as of documents, which may not be given in a number of special-purpose domains. Also, it is not clear how to make effective use of an unstructured plethora of tags, and spam blogs are beginning to threaten tag validity.

Some blog sites have recognised the need for supporting authors in their tag choice. `www.tagyu.com` suggests a blog tag for a URL or free text entered interactively or via a Web service interface. Proposals are made based on the similarity to existing, tagged blogs; this similarity appears to be computed based on syntax only (`http://www.tagyu.com/faq`).

Text similarity can also be used for discovering missing links in a corpus. For example, Adafre and de Rijke (2005) use page title information and co-citation analysis to cluster pages in Wikipedia and to identify candidate links from those similar pages that might be missing on a given page.

Blog search engines generally do a keyword search in the blogs themselves or in their RSS feeds, employing general search engine technology on these sources (e.g., `www.technorati.com/search` or `blogsearch.google.com`).

In this paper, we use text and graph mining for two purposes: First, to identify topics of a given blog corpus. This could be employed, for example, for proposing tags that – if accepted and assigned to a blog entry by a blog author – may help to increase the visibility of a blog or blog entry. Therefore, it could also be employed by blog search engines to improve keyword search.

Second, we aim to identify interconnections between blogs, i.e. find the "hidden network" induced by common topics that are so far only implicit in the textual content. This could be employed, for example, for proposing hyperlinks or blogroll items to an author, thus – if accepted by the

author – enhancing the explicit blog network. It could also be employed by blog search engines for similarity searches and recommender systems.

However, in contrast to the methods discussed above, we do not rely on syntax and statistics, i.e. the pure textual content of the blogs, alone. Instead, we consider syntax-based analysis as a baseline and show the benefits of enhancing the analysis by semantic information. In contrast to syntax-only approaches, this also allows us to improve the usability of blog-analysis applications by providing *explanations* of the suggestions made to blog authors or searchers.

The data used for the experimental analysis

The data used in our experiments consisted of blogs arbitrarily sampled from the Yahoo! blog directory categories "food and drink", "health and medicine", "law", and "weblogs about blogging"[4]. In the following, we denote the samples as *blog corpora* with the names "food" (24 Web pages with varying numbers of blog entries and a total of 330,464 words), "health" (5 pp., 251,354 w.), "law" (27 pp., 300,649 w.), and "meta-blogs" (20 pp., 141,369 w.).[5]

Reading the blogs, we found a number of common subjects in the food and health blogs, as well as in the law and meta-blog blogs. The other pairs did not seem to have overlapping content. In the following, we present an exploratory analysis of the blogs and their relations, in the tradition of, e.g., (Adamic & Glance, 2005). Towards this end, we first carried out a syntactic analysis and then a semantic analysis. These will be described in the next two sections.

Syntactic analysis

Terminology is the set of words or word strings that convey a single, possibly complex, meaning within a given community. In a sense, terminology is the surface appearance, in texts, of the domain knowledge of a community.

As a baseline characterisation, in our experiments we extracted the terminology referred within each blog corpus based on two measures, *Domain Relevance* and *Domain Consensus*, that we introduce hereafter.

High frequency in a corpus is a property observable for terminological as well as non-terminological expressions (e.g., last week or real time).[6] We measure the specificity of a terminological candidate with respect to the target domain via comparative analysis across different domains. To this end a specific score, called *Domain Relevance* (*DR*), has been defined. A quantitative definition of the *Domain Relevance* can be given according to the amount of information captured within the target corpus with respect to a larger collection of corpora. More precisely, given a set of n domains $\{\Delta_1, \ldots, \Delta_n\}$ and related corpora, the domain relevance of a term t in class Δ_k is computed as:

$$DR_{t,k} = \frac{P(t|\Delta_k)}{\max\limits_{1 \leq j \leq n} P(t|\Delta_j)} \quad (1)$$

where the conditional probabilities ($P(t|\Delta_k)$) are estimated as $E(P(t|\Delta_k)) = f_{t,k} / \sum\limits_{t' \in \Delta_k} f_{t',k}$, and $f_{t,k}$ is the frequency of t in Δ_k (i.e. in its related corpus).

Terms are single- or multi-word expressions whose meaning is agreed upon by large user communities in a given domain. A more selective analysis should take into account not only the overall occurrence of a term in the target corpus but also its appearance in single documents. Domain terms (e.g., grocery store) are used frequently throughout the blogs of a domain, while there are certain specific terms with a high frequency within single blogs, but completely absent in others.

Distributed usage expresses a form of *consensus* tied to the consolidated semantics of a term (within the target domain) as well as to its centrality in communicating domain knowledge. A second relevance indicator is therefore assigned to candidate terms, called *Domain Consensus* (*DC*). DC measures the distributed use of a term in a domain Δ_k. The distribution of a term t in blogs $b \in \Delta_k$ can be taken as a stochastic variable estimated throughout all $b \in \Delta_k$. The entropy H of this distribution expresses the degree of consensus of t in Δ_k. More precisely, the domain consensus is expressed as follows:

$$DC_{t,k} = \sum_{b \in \Delta_k} \left(P_t(b) log \frac{1}{P_t(b)} \right) \quad (2)$$

where $E(P_t(b_i)) = f_{t,i} / \sum\limits_{b_j \in \Delta_k} f_{t,j}$.

Filtering non-domain candidate terms is based on the thresholding of measures (1) and (2). General-purpose corpora previously used in different ontology learning experiments (Navigli & Velardi, 2004) were employed as contrastive corpora, and the minimal DR (DC) were set to the values 0.35 (0.23) that had proved useful in those experiments. We call the resulting terms *keyphrases* and refer to the set of keyphrases of a domain corpus C by $T(C)$.

The method, implemented in our TermExtractor tool (Navigli & Velardi, 2004), was applied to the four blog corpora introduced in the Data section. This analysis produced large numbers of keyphrases: 639 (law), 280 (food), 138 (health), and 140 (meta-blogs). This very large set of keyphrases did not lead to the emergence of a clear meaning of the corpora.

To find relations between the blog corpora, we investigated the *degree of similarity of keyphrases* for each pair of corpora C, C' by investigating their common keyphrases. We use the popular and proven Jaccard coefficient, cf. (Haveliwala et al., 2002)

$$sim(C, C') = \frac{|T(C) \cap T(C')|}{|T(C) \cup T(C')|}. \quad (3)$$

To concentrate on the most characteristic keyphrases for each corpus, we also computed sim for characteristic

[4] http://dir.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Weblogs/

[5] The blog corpora are available at http://www.wiwi.hu-berlin.de/~berendt/Blogs/Sample20050917/

[6] Throughout the article, we typeset terms and WordNet synsets in sans serif font and domain labels in SMALL CAPS. In places, their names are changed into plural form to increase readability.

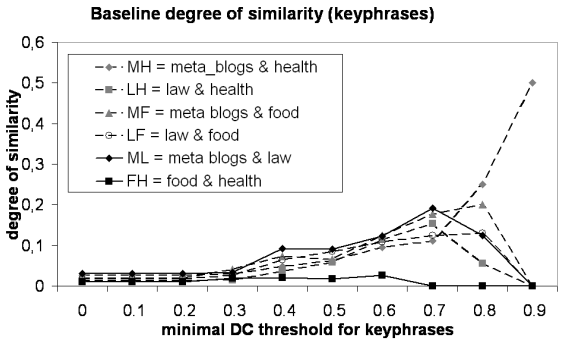

Figure 1: Syntactic similarity. Solid lines = similarity was expected, dashed lines = no similarity was expected.

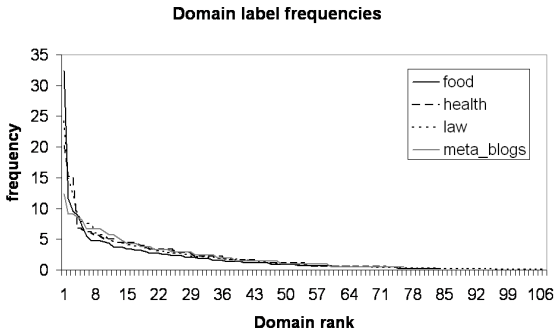

Figure 2: Distribution of domains. The lines were simplified (no markers at individual ranks) to increase readability.

keyphrases only. Characteristic keyphrases were derived by further thresholding the DC measure.[7] Figure 1 shows the degrees of similarity over all keyphrases, over all keyphrases with $DC \geq 0.1$, etc. The degree of similarity for keyphrases with $DC \geq 0.9$ was 0 in 5 of the 6 pairs and 0.5 (a statistical artifact resulting from intersection size 1 and union size 2) in 1 pair. In general, the absolute sizes of intersections and unions led us to conclude that DC filters up until 0.7 are useful and show the same relative results for all pairs, while higher DC filters lead to statistical artifacts.

The most striking result is that "law" and "meta-blogs" appear to be most closely related, while another pair that we had expected to be related (health and food) exhibits the lowest degree of similarity. All other pairs are in between.

A qualitative analysis shows that of the 23 keyphrases in the intersection of law and meta-blogs, 8 concerned blogging terminology or specific blog references (blog post, web page, movable type, front page, last page, last post, new blog, recent entry), 3 concerned legal aspects of the Internet (commons licence, creative commons licence, open source), 1 concerned other media (new book), and 1 concerned politics (national security). The remainder were general terms such as next time or great example. This is a first hint that law and meta-blog bloggers may put special emphasis on media questions including, but not limited to, the Internet. In fact, we had expected a joint interest of these two blog corpora especially in intellectual-property questions.

Health and food, on the other hand, shared just one domain-specific term (grocery store) and 3 general terms. The remaining pairs did not allow us to derive any concentration on topic groups. In summary, the syntactic analysis gave only a very dim picture of what the blogs were about.

Semantics-enhanced analyses

Enriching keyphrases by domain labels

In this section we use two resources to semantically enhance our blog analysis: WordNet (Fellbaum, 1998), a computational lexicon of English, encoding concepts in the form of synonym sets (*synsets*), and the domain labels from IRST (Magnini & Cavaglià, 2000), providing a mapping between WordNet synsets and 165 taxonomically-structured domains (e.g. DENTISTRY is a kind of MEDICINE, MEDICINE is an APPLIED SCIENCE, etc.). Notice that one term may belong to different domains, depending on the senses it denotes: For example, sense #5 of "operation" in WordNet is mapped to the domain label MEDICINE, while sense #7, "operation" in the sense of data processing, is associated with COMPUTER SCIENCE, as well as to further domains.

We used the mapping to assign a set of domains D to each corpus C, denoted as $D(C) = \bigcup_{t \in C} D(t)$, where t is a term and $D(t)$ the set of domains over all WordNet senses of t. The resulting domain-label frequency distributions over the set of terms in C are shown in Fig. 2; they indicate that in spite of the large syntactic heterogeneity of each blog collection, each corpus centres on a small number of domains.

The numbers show that, for example, 32.35% of the terms extracted from the compounds (either multi-word terms with a synset in WordNet, or the single-word constituents of the compounds) in the food blog corpus had at least one WordNet sense in the highest-ranking domain in that corpus. The 5 highest-ranking domains in the corpora were: GASTRONOMY, ALIMENTATION, QUALITY, BOTANY, and PERSON (food); MEDICINE, TIME PERIOD, QUALITY, BIOLOGY, and PHYSICS (health); LAW, QUALITY, POLITICS, ADMINISTRATION, and ECONOMY (law); TELECOMMUNICATIONS, TIME PERIOD, PERSON, PUBLISHING, and ECONOMY (meta-blogs). The first three could be expected; they are more or less the same concepts that the Yahoo! labellers chose to group the blogs in their directory. (The result thus also validates the choice of the domain labels from IRST to summarise blog content at a very coarse level). The fourth differs in an interesting way: The rather young concept of meta-blogging is not a domain label in long-established thesauri / category systems. However, the top-ranking domain labels each describe an important aspect of meta-blogging (with the possible exception of TIME PERIOD and PEOPLE, see below).[8]

[7] Alternatively (and with similar results), one can compute similarity for the top n, $n \geq 1$, keyphrases of each pair.

[8] Blog exists both as a noun and as a verb in WordNet, it is a direct hyponym of diary/journal, which in turn is a hyponym of writing – indeed making PUBLISHING a good description.

The results show two further things: First, in spite of the identification of highly descriptive domains by the top-ranking domain labels, generic expressions constitute a large part of the text, coming from domains such as TIME PERIOD, PERSON, or QUALITY. Magnini and Cavaglià (2000) group these domains under the generic super-domain FACTOTUM. Second, while the top-ranking domain labels describe each corpus well *in isolation*, they fail to capture some important *relations* between corpora. This was investigated by studying the intersections between top domains. In the following, $D^n(C)$ is the set of the n top-ranking domains from $D(C)$.

The top domains were largely disjoint: there were 0 common top domains (intersection of the $D^1(C)$). In the top 3 domains $D^3(C)$, food / health, food / law, law / health, and health / meta-blogs each shared 1 domain, and in each case, this was a factotum domain (QUALITY, TIME PERIOD). These were also the intersections of the top 5 domains $D^5(C)$; in addition, food and meta now shared the factotum domain PERSON, and law and meta shared the meaningful domain ECONOMY. In the top-10 domains $D^{10}(C)$, law and meta-blogs were most related both quantitatively and qualitatively: They shared 6 of their top-10 domain-labels, and 3 of these were meaningful (LAW, POLITICS, ECONOMY); the remainder were factotum domains. Law and health ranked second, with 2 of 4 meaningful (LAW, PSYCHOLOGY), and health and meta-blogs both dealt with LAW issues (and 2 factotum concepts). Food and meta-blogs, and food and law only shared small numbers of factotum concepts. Food and law also shared SOCIOLOGY, which in the case of food was however due to very generic terms which just happened to be mapped to sociology (world,place,commons,fame,family,national,show,public). Surprisingly, there was still no overlap between the domain labels implicated in food and health.

This indicates that taxonomic domain knowledge alone is not sufficient to capture the meaning of blogs and their interrelations.

Domain labels and Structural Semantic Interconnections

The main point of the experiment described in this section was to investigate non-taxonomic relations between keyphrases in order to obtain a broader semantic picture of the meaning, and the common meaning, of blog corpora. For this purpose we employed the SSI algorithm (Navigli & Velardi, 2005). SSI (Structural Semantic Interconnections)[9] is a knowledge-based algorithm for Word Sense Disambiguation. Given a word context and a lexical knowledge base, obtained by integrating WordNet with annotated corpora and collocation resources, SSI selects a semantic graph including those word senses having a higher degree of interconnection, according to a measure of connectivity based on the number and weight of semantic interconnection patterns.
A *semantic interconnection pattern* is a relevant sequence of edges selected according to a context-free grammar, i.e. a path connecting a pair of word senses, possibly including a number of intermediate concepts. In the following, we therefore also refer to these patterns as "paths".

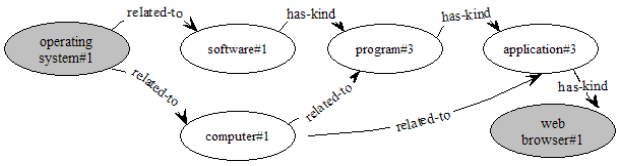

Figure 3: Examples of SSI interconnections.

Some example paths are shown in Fig. 3, showing, among others, that an operating system is related to[10] software, that programs are a special kind of software (hypernym relation), applications a special kind of programs, and web browsers a special kind of applications. The meanings of the terms along the paths are indexed by their WordNet sense keys. Both target and source are mapped to the domain label COMPUTER SCIENCE. The weight of this path is 0.25, which is 1 divided by the number of edges.

In addition, we used the taxonomy of the domain labels introduced in the previous section. For example, both MEDICINE and COMPUTER SCIENCE are subcategories of APPLIED SCIENCE, and MEDICINE is the supercategory of DENTISTRY, PSYCHIATRY, and others.

The following notation is used: $D(C)$ is, as defined above, the set of all domains associated with corpus C, and $D^n(C)$ contains the top n domains.

$\mathcal{D}$ is the set of abstractions of a set of domain-labels D (the labels in D and all their ancestors in the domain-label taxonomy). $\mathcal{D}^n$ is defined analogously.

$D^n_+(C)$ is the set of domains that are *related to* at least one domain in $D^n(C)$. This *domain relatedness* is defined as follows: domain D' is related to D if $|\{(s,s') \in Synsets \times Synsets : s$ is mapped to the D domain, s' is mapped to the D' domain and $s \stackrel{related-to}{\longrightarrow} s'$ in the SSI lexical knowledge base $\}| \geq \theta$, where $Synsets$ is the set of WordNet concepts, and θ is a threshold experimentally fixed at 0.02. In other words, a domain D' is related to D if a certain number of concepts mapped to D' are related to concepts mapped to D through an edge in the lexical knowledge base. For example, while COMPUTER SCIENCE and TELECOMMUNICATIONS are in different branches of the domain-label taxonomy, they are strongly related. $\mathcal{D}^n_+(C)$ is defined analogously.

We first established a baseline that used no domain information, but only the semantic disambiguation inherent in SSI: Based on the assumption that multi-word expressions (*compounds*) are highly characteristic of a domain, we derived all the semantic interconnections between the compounds of each pair that were no longer than 4 edges. This is similar to the syntactic baseline established above; however, whereas there we investigated compounds that were *included* in both corpora of a pair, here we investigated compounds that were *related* by being the source and target of a semantic interconnection.

We then investigated the usefulness of adding domain information, filtering the source and target expressions of semantic interconnections by varying the following factors:

[9] SSI can be accessed online at: http://lcl.di.uniroma1.it/ssi

[10] The "related-to" relation is derived from semantically disambiguated collocations (Navigli, 2005).

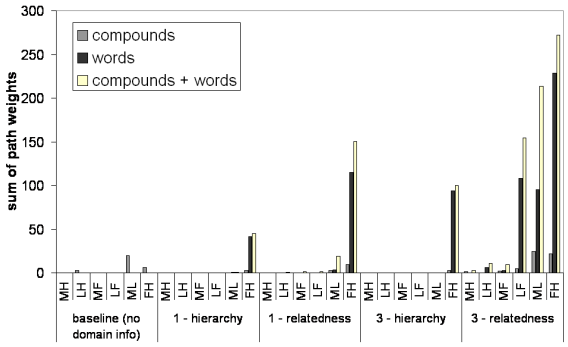

Figure 4: Semantic interconnectedness.

A. Choice of terms

compounds : characteristic compounds as identified in the baseline setting above that have a WordNet sense and thus also a domain label (e.g. hard disk with the domain COMPUTER SCIENCE)

words : all single-word terms that are part of the compounds

compounds + words : both

B. Specificity

1 : only domain labels from $\mathcal{D}^{\,1}_{(+)}(C)$ are considered

3 : only domain labels from $\mathcal{D}^{\,3}_{(+)}(C)$ are considered

Specificity acts as a filter on the terms: only those that map to a domain label in the set $\mathcal{D}^{\,n}_{(+)}$ are considered as source or target of a semantic interconnection. $\mathcal{D}^{\,n}_{(+)}$ is short for $\mathcal{D}^{\,n}$ or $\mathcal{D}^{\,n}_{+}$.

C. Structure

hierarchy : only the domain-label taxonomy is used for choosing $\mathcal{D}$, i.e. all terms that have at least one sense in $\mathcal{D}^{\,n}(C)$ can be source or target of a semantic interconnection

relatedness : taxonomy and domain relatedness are used, i.e. all terms that have at least one sense in $\mathcal{D}^{\,n}_{+}(C)$ are taken as source or target of a semantic interconnection.

In other words, we investigated the effects of two kinds of semantic enrichment: First, whether an analysis of the terms in a corpus can profit from a focus on the most frequent domains and their super-domains or not ("specificity"), and second, whether the analysis can profit from a thematic focus *plus* a guided enlargement of this focus by non-taxonomic relations between domains ("structure").

All 3*2*2=12 combinations of factor levels were investigated, and they were compared to the pure-SSI baseline.

As a dependent variable, we collected all paths returned by SSI between all source and target terms in senses that mapped to the domain labels as specified by the respective experimental condition's specificity+structure. We calculated the total weight of a connection between two blog corpora as the sum of the single path weights.

Results: Quantitative interpretation. As a summary of the results, we plot the path value sum in Fig. 4, where the value of a single path is 1 divided by the number of its edges. The figure shows a clear pattern of main effects and interactions: (a) Both the relaxation of the restriction in numbers of domains (factor "specificity") and the integration of non-taxonomic semantic relations (factor "structure") increase interrelatedness. However, when both relaxations are combined, nearly all blog pairs appear related. The order of relatedness, conforms with our initial expectations: Food and health appear highly related, followed by meta blogs and law. In the "3-relatedness" conditions, there is in addition a relation between food and law. (b) The combination of syntax and semantics seems to have an additive effect on the results; there is no marked difference between the quantitative patterns at the three levels of factor "choice of terms".

Results: Qualitative interpretation. An analysis of the paths in the pure-SSI baseline discovered, for law / meta-blogs, a large number of interconnections through terms associated with COMPUTER SCIENCE and, a distinct second, PUBLISHING. Thus, while computer science terms are themselves not highly characteristic of either law or meta-blogs[11], they constitute a highly interconnected common core of these two corpora and might therefore constitute a reason for recommending one to readers and authors of the other. In the following paragraphs, we provide a qualitative interpretation of the results for the three different choices of terms (compounds, single words, compounds + words).

Compounds However, while the connection itself may be correct, it is so for the wrong reasons. A recommendation based on 'hard drives' as a joint topic of law and meta-blog blogs does not appear to be very convincing. A second, but much weaker, group of paths is more likely to catch such readers' true intentions. These paths deal with PUBLISHING, linking reference books and news stories.

The food–health pair has a number of paths explaining different relations between types of fat or greasy food (cream cheese, chocolate sauce, fatty acid, vegetable oil, ice cream, hot dog) to other fats, or to health food.

The remaining paths in this condition were factotum links between law and health (based on the term long time).

Filtering the compounds by whether they mapped to senses in the top domains retained the food–health paths, but eliminated all other paths. No compound-based paths were left between law and meta blogs because computers are the main domain of neither of them. A relaxation to the top 3 domains did not affect this pattern of interconnections.

When related domains are also taken into account, the spurious computer-induced relation between meta-blogs and law does not re-appear because while computer science is related to telecommunications, it is not related to law. However, interesting new connections appear: The law-related domain ENTERPRISE (which mirrors the common domain ECONOMY, see Section on enriching keyphrases by domain labels) has many connections with the TELECOMMUNICATIONS-related domain COMPUTER SCIENCE, dealing with such contents as law firms or news organizations. In addition, there are publishing-related connections

[11]The domain LAW ranked 8th in meta-blogs, and the domain COMPUTER SCIENCE ranked 7th (both with a frequency of 6.67). In the law corpus, PUBLISHING ranked 14th (4.48), TELECOMMUNICATIONS 25th (2.78), and COMPUTER SCIENCE 32nd (2.15).

from news organization (ENTERPRISE) to news story (PUBLISHING) and to political party (POLITICS).

Examples include the following 4 paths: A law firm#1 (domain ENTERPRISE) is-a firm#1, which is-a business#1. Business is-related-to computer#1, which is related to domain name#1, operating system#1, computer networks#1, and source code#1 (all in the domain COMPUTER SCIENCE). Another path specifies that a news story#1 is-an article#1, which is related to a report#2, which in turn in related to news organization#1.

In the food–health pair, the hierarchy-based paths are enhanced by some connections based on PHYSICS (linking body weight and body temperature to various food items). Also, the path set is enlarged by semantically similar paths: Because MEDICINE and GASTRONOMY are themselves related, this condition adds MEDICINE → GASTRONOMY paths to the previously derived GASTRONOMY → MEDICINE paths (e.g., medical side effects are discussed in the food blog corpus, while food products which may cause them are discussed in the health blog corpus). This is a particularly interesting pattern: topics from domain group 1 are discussed in domain group 2, and vice versa. We refer to this pattern as the *topic-reversal effect*.

No further paths emerge in this condition.

The relaxation to the top 3 domains plus relatedness makes some previously lost connections re-appear, in particular, the COMPUTER SCIENCE connections between meta-blogs and law. In addition, some potentially useful connections based on media appear, but both their rarity and their content suggest a spurious connection (news organization – black-and-white photography).

Unexpected but potentially useful connections emerge in the law–food pair: local government is linked to town planning, including parking lots and the main drag (the main street with grocery stores and other shops). These paths also occur in the food–health pair, suggesting the influence of town-planning decisions on nutrition and, consequently, health as a joint blog topic.

The remaining paths are artifacts (meta-blogs–food: bank accounts and phone numbers, which may be general-purpose contact info on Web pages, are linked to different items of POLITICS and PUBLISHING). An indication of their poor quality is the very small number of concepts that are linked by many different interconnections.

Single-term expressions ("words") In the basic condition (1-hierarchy), an interesting pattern is observed: The food–health interconnections are enhanced substantially. Various items of food (eggs, onions, bacon, etc.) and apparel/personnel (cook, dish) are all related to health food, along different paths. In addition, disease is linked to beef. All paths are meaningful. The substantial difference to the compounds condition may be interpreted as showing a large extent of lexicalization in the discourse domains food and health, which are very "basic" to human life. We call the resulting increase in interconnections between single words (as opposed to compounds) the *lexicalization effect*.

When related domains are also taken into account (1-relatedness), a similar phenomenon is observed between meta-blogs and law: Various media or publication forms, all of which have single-term names (medium, video, radio, tv show) are now linked to key concepts in law: copyright, damages, security, penalty, and lawyer.

The same pattern, combined with the topic-reversal effect, increases the number of paths between food and health. In addition, drinking (a term which is classified as PHYSIOLOGY) is now related to health questions. An artifact seems to be the frequent occurrence of section, a term which is classified into MEDICINE and linked to disease, but which (in the food blog) probably refers to cutting food.

The relaxation to 3 top domains has no effect.

The extension to 3 top domains plus relatedness appears to be only partially helpful, because a large number of very general paths appear: books (meta-blogs) are published on every conceivable topic in law and its related domains, marketing (meta-blogs) can be applied to every item of everyday use (food), and cities (law) can cause diseases (health).

Highly generic single-word terms (activity, life, computer, area, food, kind in food) establish a large number of generic connections to arbitrary terms from a second corpus, because these terms are "related" to nearly everything else. We call the large increase in the number of paths that involve generic terms the *topic drift effect*.

In the law–food pair, many intra-domain paths (gastronomy–gastronomy) are found (the *identity effect*).

The only meaningful new connections in this condition are COMMERCE–POLITICS connections in meta-blogs–law.

Compounds + words When multi-word and single-word terms are combined, the paths are just put together. However, this aggregation may help to alleviate strong lexicalization effects in one pair by counteracting them with topic-reversal effects in another pair. In the present case, however, the order of pairs was very distinct in both the compounds and the words conditions; the order therefore did not change in the compounds + words conditions.

Restricting path grammar

Starting from the 3-relatedness condition, to counteract the topic drift effect, we restricted the grammar of admissible semantic interconnections. First, we removed paths that involved two "related-to" links. This filtered out 88.8% of the paths. Then, we removed paths that involved more than two types of links. This filtered out 53.4% of the paths.

Only three corpus pairs remained that were classified as having valid interconnections (in the order of path weight sums): food–health, meta-blogs–law, and law–food. The final paths may be used to derive explanations of the recommendation of one blog corpus as related to the other one.

It should be noted that the full effect of this way of post-processing paths, in terms of false positives and false negatives, remains to be investigated more fully: While many immediately useless-looking paths were pruned, pruning also sacrificed some paths involving two "related-to" links that were meaningful (cf. the example given with respect to the "compounds" conditions).

Summary and general discussion of the results

In this paper, we have proposed the use of semantics to improve the analysis of blogs. We investigated topic analysis of a blog or set of blogs and interrelations between blogs. Applications include the proposal of tags for annotating blogs, the recommendation of "also interesting" blogs, and the improvement of keyword search or similarity search.

We showed that based on a simple syntactic analysis of common keyphrases, some connections could be found, but that this may have been a chance finding: while two blogs that were manually identified as touching on similar issues (law and meta-blogs) were also similar on the surface level, a second similar pair (food and health) scored lowest.

We then enhanced the analysis by focusing on interrelations between terms in each pair, induced by their associated concepts in WordNet and additional information encoded in the SSI (Structural Semantic Interconnections) tool. While this produced – quantitatively – the right pairs, the qualitative information was poor, such that the findings could again be due to chance and extremely sensitive to the sample used.

Third, we mapped significant terms from the blog corpora to their domains, and we used a general-purpose domain taxonomy and (non-taxonomical) domain-relatedness information to filter the meaningful paths from the full set of SSI interconnections. This made the quantitative results converge on the expectations based on the manual inspection of the corpora, and the semantic paths which constitute the qualitative results became more meaningful. In particular, domain relatedness proved to be useful.

We identified syntactical-semantical patterns in these interconnections that we summarized as four effects. Their understanding can guide processing to find the most interesting and valid interconnections.

Fourth, to reduce topic drift, we suggested restrictions on path grammar. The quantitative and qualitative results improved. In summary, we can conclude that the integration of domain semantics and of structural semantics is an invaluable help in the analysis of blogs and their interconnections. In particular, meaning and interconnections are not only found, they can also be explained.

Conclusions and outlook

This work is just a first step towards a semantic understanding of blogs. In future work, we plan to investigate more fine-grained semantic relations (including a graded notion of domain relatedness), and to standardise the data preprocessing based on contrast corpora. In addition to this, we plan to extend our work with the general-purpose ontology WordNet to more domain-specific ontologies, as well as to apply our tools for ontology learning to enrich WordNet by newly emerging topics of discourse that appear in blogs.

In follow-up studies, we aim at using larger samples (this would also allow us to investigate the statistical significance of results), and to compare our work with other blog analysis techniques and software, in particular with approaches that rely on syntax (e.g., Gruhl et al., 2005) and/or explicit user ratings (e.g., Tremblay-Beaumont & Aïmeur, 2005) as a baseline. If samples with rich manual tagging become available, it will moreover become possible to assess the different methods with respect to the precision and recall of topic and relatedness detection. Corpora such as the WWW 2006 Weblogging Ecosystems Workshop data challenge[12] mark a first step in this direction. However, these data are tagged in folksonomy style, which is radically different from the "classical" tagging needed for standard recall/precision analyses. Thus, the development of adequate methods for processing folksonomy tags is another relevant research direction (clustering may be a promising approach, cf. `www.flickr.com`).

A second essential issue is scalability. Our methods - like other forms of exploratory analysis, such as (Adamic & Glance, 2005) - scale quantitatively, but the qualitative judgment of the result of course needs human input. In future work, we intend to explore mass collaboration as a means of making the second step scalable.

References

Adafre, S.F., & de Rijke, M. (2005). Discovering missing links in Wikipedia. In *Proc. of the* 3^{rd} *Int. Worksh. on Link Discovery* at *ACM SIGKDD* (pp. 90–97).

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. Election: Divided they blog. In *Proc. of the* 3^{rd} *Int. Worksh. on Link Discovery* at *ACM SIGKDD* (pp. 36–44).

Adar, E. Glance, N., & Hurst, M. (Eds.) (2005). *2nd Annual Workshop on the Weblogging Ecosystem* at *WWW 2005*. `http://www.blogpulse.-com/www2005-workshop-cfp.html`[13]

Borgs, C., Hayes, J.T., Mahdian, M., & Saberi, A. (2004). Exploring the community structure of newsgroups. In *Proc.* 10^{th} *ACM SIGKDD* (pp. 783–787).

Fellbaum, C. (Ed.) (1998). *WordNet: an Electronic Lexical Database*. MIT Press.

Glance, N., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *Workshop on the Weblogging Ecosystem* at *WWW 2004*. `http://www.blogpulse.com/papers/www2004glance.pdf`

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005). Deriving Marketing Intelligence from Online Discussion. In *Proc.* 11^{th} *ACM SIGKDD* (pp. 419–428).

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The Predictive Power of Online Chatter. In *Proc.* 11^{th} *ACM SIGKDD* (pp. 78–87).

Haveliwala, T., Gionis, A., Klein, D., & Indyk, P. (2002). Evaluating strategies for similarity search on the web. In *Proc. of WWW 2002* (pp. 432–442).

Magnini, B. & Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In *Proc. LREC-2000* (pp. 1413–1418).

Navigli, R. (2005). Semi-automatic extension of large-scale linguistic knowledge bases. In *Proc. of 18th FLAIRS* (pp. 548–553).

Navigli, R. & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Websites, *Computational Linguistics*, 30(2), 151–179.

Navigli, R. & Velardi, P. (2005). Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7), 1075–1086.

Quintarelli, E. (2005). Folksonomies: power to the people. *ISKO Italy-UniMIB meeting.* `http://www.iskoi.org/doc/folksonomies.htm`

Tremblay-Beaumont, H., & Aïmeur, E. (2005). Feature combination in a recommender system using distributed items: The case of JukeBlog. In *Proc. of Multi-Agent Inform. Retrieval and Recommender Systems WS* at IJCAI-05 (pp. 70–74).

Yi, J. (2005). Detecting buzz from time-sequenced document streams. In *Proc. of IEEE EEE 2005* (pp. 347-352).

[12] `www.blogpulse.com/www2006-workshop/#data`

[13] Access date of all online references: 27 January 2006