Language Model Mixtures for Contextual Ad Placement in Personal Blogs

Gilad Mishne and Maarten de Rijke

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
`gilad,mdr@science.uva.nl`

Abstract. We introduce a method for content-based advertisement selection for personal blog pages, based on combining multiple representations of the blog. The core idea behind the method is that personal blogs represent individuals, whose interests can be modeled by the language used in the blog itself combined with the language used in related sources of information, such as comments posted to a blog post or the blogger's community. An evaluation of our ad placement method shows improvement over state-of-the-art ad placement methods which were not designed for blog pages.

1   Introduction

Blogs—frequently modified web pages in which dated entries are listed in reverse chronological order—come in a variety of genres [1]. In this paper, our focus is on personal blogs, created by individuals and serving as a vehicle for self-expression and self-empowerment; this type of blogs is by far the most common. Personal blog posts are often not topically focused—instead, they provide reports about experiences and interests of individuals, and of the objects they surround themselves with and the activities they engage in. This is one of the major differences between the text found in typical personal blog posts, and the text found in other web pages: whereas most web pages represent information, personal blogs represent individuals.

Blogs and the blogosphere form an increasingly active area of research, with interest ranging from language technology and text mining to information access, as is witnessed by e.g., the launch of a blog track at TREC 2006 [2]. Alongside the academic interest, blogs are also a rich source of information for commercial purposes. At the aggregate level, various uses have been made of blogs and their contents, e.g., predicting spikes in consumer purchase decisions using the mere volume of blog postings [3]; at the individual blog level, tools such as book recommender systems based on bloggers' writings have been proposed [4]. In this paper, we are interested in developing language technology for a different commercial aspect of blogs: advertisement placement. Specifically, we want to generate suggestions for advertisements to be displayed to readers of a blog, based on the content they are viewing. This type of ad placement is sometimes
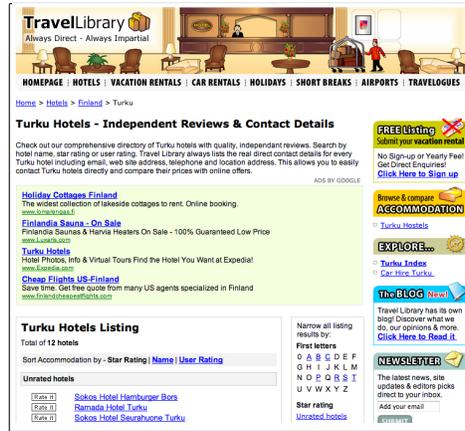
Fig. 1. Contextual ads on a non-blog page.

called *contextual* or *content-based*, since the displayed ads are related to the context in which they appear. For example, if the reader is viewing a blog post discussing sports, and the blogger's site uses contextual advertising, the user might see ads from advertisers such as sports memorabilia dealers or ticket sellers. Figure 1 shows an example of contextual ad placement in a non-blog page: in this example, Google's AdSense program selects ads to display in a website reviewing hotels in Turku, Finland (ads appear in the central part of the page).

Briefly, then, the task we address is this: given a personal blog post and a collection of advertisements, identify advertisements that are most relevant for the post: advertisements that are most likely to be of interest to readers of the post.

Our research is driven by two main questions. First, contextual ad placement methods have been developed for general (non-blog) web pages. How do these methods perform on personal blogs (as opposed to non-blog web pages)? We find that ad placement in personal blogs is harder than in other web pages: a state-of-the-art method developed for web pages does not achieve the same results on personal blogs. One of the truly challenging aspects of personal blogs for contextual ad placement is that personal blogs tend to be non-topical, meaning that there is no "real" topic to many of their posts—a real problem for ad placement methods that rely on identifying the general topic of a page. This observation motivates our proposal of an alternative placement algorithm, one that takes the view that a personal blog represents a person, not a topic, as its starting point. Our second research question, then, is whether this person-oriented approach yields a more effective ad placement method for personal blogs than state-of-the-art placement methods built for generic web pages.

The rest of the paper is organized as follows. In Section 2 we discuss related work. Our ad placement approach, based on person-oriented language model mixtures is presented in Section 3. Section 4 contains a description of our experimental evaluation; we conclude in Section 5.

2 Related Work

First deployed in 2003, contextual ad placement services allow websites to pay to have their advertisements displayed alongside the contents of related web pages. Programs such as Google's AdSense and Yahoo's Publisher Network are effective in generating revenue both for the advertiser and the ad-matching mediator by associating the content of a web page with the content of the displayed ads, increasing the likelihood of their usefulness. Often, the ads are non-intrusive and are clearly marked as such; on top of that, they enjoy the reputation of the ad selection platform (which is typically a well-known web search engine)—this explains much of the success of contextual ad placement [5].

As contextual ad placement has become a substantial source of revenue supporting the web today, investments in this task, and more generally, in the quality of ad placement, are increasingly important. Most of the advertisements are currently placed by search engines; advertisements that are not relevant may negatively impact the search engine's credibility and, ultimately, market share [6,7]. The more targeted the advertising, the more effective it is [8]. As a consequence, there has been a considerable amount of research on relevance in advertising for general web data (see Section 2).

As the area of content-based ad placement is relatively new, and since it involves many "trade secrets," the amount of existing published work is limited. The work most closely related to ours is that of Ribeiro-Neto et al. [9], involving an impedance coupling technique for contextual ad placement. This approach uses a variety of information sources, including the text of the advertisements, the destination web page of the ad, and the triggering words tied to a particular ad. We use the AAK_EXP method described in Ribeiro-Neto et al.'s work as state-of-the-art, for comparing with our approach. Work on ad placement prior to [9] was of a more restricted nature. E.g., [10] propose a system that is able to adapt online advertisements to a user's short-term interests; it does not directly use the content of the page viewed by the user, but relies on search keywords supplied by the user to search engines and on the URL of the page requested by the user. Finally, [11] report not on matching advertisements to web pages, but on the related task of extracting keywords from web pages for advertisement targeting. The authors use various features, ranging from *tf* and *idf* scores of potential keywords to frequency information from search engine log files.

3 Language Model-Based Blogger Profiles

Contextual placement of text advertisements boils down to matching the text of the ad to the information supplied in a web page. Typically, a textual ad is composed of a few components: the self-explanatory *title*, designed to capture the attention of the viewer, a short *description* providing additional details, a *URL*, the target a surfer will be taken to if the ad is clicked, and a set of *triggering terms*. The triggering terms, which are not displayed to the surfer, are provided by the advertisers and function as terms associated with the ads, assisting the process of matching ads with context. In this paper we follow a

standard approach which concatenates the text of all these different components to a single textual representation of the advertisement. The challenge we are facing is to select ads (from a collection of ads represented in this concatenated manner) that are most likely to be of interest to readers of a given blog post.

As outlined earlier, our working hypothesis is that personal blogs represent individuals. In this section we develop a framework for modeling these individuals using statistical language models, and matching these models to the advertisements. First, we provide some background about language models; we follow with an instantiation of these models for blogs.

3.1 Language Models and Model Similarity

Language models are statistical models that attempt to capture regularities of natural language phenomena [12]. Long in use by the speech recognition community, in the last decade they have been successfully adopted by researchers in other areas such as information retrieval [13] and machine translation [14].

The language models we use are probability distributions over sets of strings, where the probability assigned to a string is the likelihood of generating it by a given language. To estimate the probabilities, we use a maximum likelihood estimate generated from observed text. We use the most common type of language model: unigram models, in which the strings are single-word terms from the language's vocabulary. In practice, then, our language models consist of probabilities assigned to words according to their frequency in the text.

Since language models are probability distributions, statistical methods for comparing distributions can be used to compare them. Applying goodness-of-fit tests to two language models—one functioning as the expected distribution and the other as the observed one—indicates to what degree they differ. While a number of such tests exist, comparisons of models of the type we use is best performed by a *log likelihood* test, since the text contains a large amount of rare events [15]. This test assigns every word in the language a divergence value indicating how different its likelihood is between the two languages: words with high log likelihood values are more typical of one language than the other, and words with low values tend to be observed in both languages with similar rates.

3.2 Information Profiles

Divergence between language models provides an elegant way of building an "information profile" of a given document (or set of documents) taken from a larger collection. First, language models are estimated both for the given document and for the entire collection. Then, these two models are compared. Ordering the terms of the models according to the divergence values assigned to them functions as the profile of the document. Prominent terms in the profile—terms with higher divergence values—are more "indicative" of the content of the document, as their usage in it is higher than their usage in the rest of the documents. For example, according to this method the most indicative terms for this paper (when compared to a large collection of other scientific articles in various computer science areas) are "blog," "model," "advertisement," and "language."

3.3 Language Models of Blog Posts

Our approach to constructing profiles of blog posts is based on forming information profiles from text as just outlined. But what "text" should we use for building this profile? In the context of a specific blog post there are different sources of information about a blogger. Clearly, the blog post itself is an important source of information. Another obvious source of information is the contents of other blog posts written by the same blogger—i.e., the contents of the blog as a whole. Some properties of blogs, such as their community-oriented structure or their temporal nature, provide additional sources of knowledge. Our approach, then, attempts to distill a textual model of the blogger by combining the information present in each of these representations.

Exploiting various subsets of the information sources listed above, we build the following models for a blog post p.

Post Model. For this model we use the most straightforward content: the contents of the blog post p itself.

Blog Model. This model is built from all posts from the same blog as p which are dated earlier than p. The intuition behind this is that interests and characteristics of a blogger are likely to recur over multiple posts, so even if p itself is sparse, they can be picked up from other writings of the blogger.

Comment Model. One of the distinct properties of blogs is the ability of blog visitors to respond directly to a post by leaving a comment which is made public on the post page [16]; these are often identified as important for the blogging experience (e.g., [17]). Our comment model is constructed from all comments posted in response to p, and assumes that their content is directly related to the post.

Category Model. *Tags*—short textual labels that many bloggers use to categorize their posts [18]—are another feature often occurring in blogs. These labels range from high-level topics ("sport," "politics") to very specific ones ("Larry's birthday," "Lord of the Rings"). For this model, we used all blog posts filed under the same category as p, as the bloggers themselves decided that they are topically related.

Community Model. Given our assumption that blogs represent individuals, it is natural that the blogspace provides fertile ground for the formation and interaction of a large number of communities [19]. This model exploits this aspect of blogs by using all text of blogs which are part of the same community as the blog p is taken from. A formal definition of a blog community does not exist; in this work, we take a simple approach and mark a blog as belonging to the community of p's blog if it links at least twice to that blog.

Similar Post Model. For this model, we use the contents of the 50 blog posts which are most similar to p. To measure similarity, we used the language modeling approach described in [20]; in practice, we indexed the entire collection of blog posts and used a language modeling-based IR engine to retrieve the top 50 posts from the collection, using p itself as a query. This model attempts to overcome the vocabulary gap which exists between some

of the relevant ads and the posts by adding terms from related posts, in a similar manner to that proposed in [9].

Time Model. Personal blogs function as online diaries. As such, many blog posts contain references to ongoing events at the time of publication. For example, the time-span of the blogs in our collection (see Section 4) includes New Year's Day 2006, with many references to fireworks and parties from various blogs around that day. To accommodate this, we construct a model based on all blog posts published in a 4-hour window around the publication time of p, capturing events that influence a large number of bloggers.

Each one of these models provides a weighted list of terms, where the weight assigned to a term is its divergence value when comparing the text used for the model with the entire collection of blogs.

3.4 Model Mixtures

Forming combinations of different language models is a common technique when applying these models to real-life tasks. While finding the optimal mixture is a complex task [21], there are methods of estimating good mixtures [21,22]. In our case, we are not combining pure language models, but rather lists of terms derived from language models. As with most model mixtures, we take a linear combination approach: the combined weight of a term t is $w_t = \sum_j \lambda_j \cdot w_j(t)$, where λ_j is the weight assigned to model j and $w_j(t)$ is the weight assigned to the term t by model j. To estimate the model weights λ_j, we combine static and on-line methods as detailed below.

Static weights. Clearly, the contribution of each of the models is not equal a-priori; for example, the model representing the blogger herself is arguably more important than the one representing the community. Optimal prior weights can be estimated for each model in the presence of training material; these constitute static weights, as they do not depend on a specific set of models derived from a given blog. In the absence of training material, we used a simple prior weighting scheme where all models have the same weight w, except the post model which gets a weight of $2w$ and the time model which gets $0.5w$—we mark this as λ_j^s.

On-line weights. In addition to the static model weights, we use posterior weights associated with a specific set of models; this type of weights is also called "on-line" [22] since they are calculated on the fly, once models have been induced by a given post. These weights are aimed at capturing the relative importance each model should have, compared to other models induced by the same blog post. In our setup, we associate this importance with the quality of the model—better formed models should have a higher weight. As detailed above, our models consist of lists of terms; one way to evaluate the quality of such a list is to check its coherency—the degree to which the different terms in the list are related (this idea is often used when evaluating textual clustering methods).

To measure this coherency, we need to estimate how related the different words in the list are. For this, we calculate the pointwise mutual information

(PMI)—the statistical dependence—between any two terms in the list, and take the mean of these values as the coherence of the list. PMI values themselves are calculated using a method called PMI-IR which employs joint and independent counts of the two terms in a large corpus [23], which is in our case a collection of blog posts. The on-line weights obtained this way are denoted as λ_j^o.

The final weight λ_j assigned to model j is $\lambda_j = \lambda_j^s \cdot \lambda_j^o$. Note that words may appear in multiple models, boosting their final weight in the combined model.

3.5 Ad Matching

Having built a combined model for blog posts, we proceed to the final phase, where the advertisements are matched to this model.

As in [9], we take an information retrieval approach to this task. Similarity between an advertisement and a model is measured with an information retrieval ranking formula—in our case, a state-of-the-art language modeling-based one [20]. We index all ads, and "retrieve" the most similar ones using a query which contains the top terms appearing in the combined divergence model described above.

Summing up, the ad selection process for a blog post p proceeds as follows.

1. Construct the different language models relating to various aspects of p.
2. Calculate divergence values for the terms in each model, when compared to a model of a large collection of blog posts.
3. Combine the diverging terms to a single weighted list using a linear combination, with a combination of static and on-line weights.
4. Use a query consisting of the top terms in the combined model to rank all advertisements; top-ranking ads are shown to the user.

In terms of complexity, the performance of our method is similar to AAK_EXP: the most demanding phase is the retrieval of additional posts for constructing the "similar-post" model, and this is done once per blog post. The background language model is static and does not require computation per post, and inducing and comparing the rest of the models is a relatively cheap process, compared with retrieval.

3.6 A Worked Example

In Table 1 we summarize the kind of information used and generated during the ad placement process, when used with a given post from our corpus[1] (for details on the corpus, see Section 4). The blog post itself deals with birds visiting the blogger's garden, and this is reflected in the post model. Additional models, in particular the community and category ones, expand the profile, showing that the blogger's interests (and, hence, the interests of visitors to the blog) can be generalized to nature and related areas.

---

[1] All our data in this paper, including the examples, is in Dutch; the examples are translated into English for convenience.

Permalink	`http://alchemilla.web-log.nl/log/4549331`
Date	January 4th, 2006
Post	*Life in the Garden* Birds are flying around the tree and the garden behind our house... Hopping blackbirds, a few red-breasts, some fierce starlings and, surprisingly, a few Flemish jays. I thought Flemish jays live in the forest. I haven't heard the trouble-making magpies from the neighbors for a couple of days, they must have joined the neighbors for their winter vacation :) I see now . . .
Post terms	garden, spot, starlings, blackbirds, (feeding)-balls
Blog terms	nature, bird, moon, black, hats, singing, fly, area
Comment terms	jays, hydra
Category terms	bird, moon, arise, daily
Community terms	nursery, ant, music, help, load, care
Similar-post terms	birds, garden, jays, blackbirds, Flemish, red-breasts
Time terms	(none)
Model weights	Post:0.63, Blog:0.21, Comment:0.02, Category:0.05, Similar-posts:0.09 Time:0
Combined model	birds, spot, garden, jays, blackbirds, nature . . .
Selected ads	www.stepstone.nl: Interested in working in nature protection and environment? Click on StepStone. www.directplant.nl: Directplant.nl delivers direct from the nursury. This means good quality for a low price. www.ebay.nl: eBay - the worldwide marketplace for buying and selling furniture and decorations for your pets and your garden.

Table 1. Example of model mixtures for ad-matching.

4   Evaluation

In this section we describe the experiments conducted to evaluate our ad placement method and the results obtained.

4.1   Experimental Setting

Blog Corpus. We obtained a collection of 367,000 blog posts from 36,000 different blogs, all hosted by web-log.nl, the largest Dutch blogging platform. The vast majority of web-log.nl blogs are diary-like, and belong to the "personal journal" blog type [1]; their content is similar to that of LiveJournal or Xanga blogs. The collection consists of all entries posted to web-log.nl blogs during the first 6 weeks of 2006, and contains 64M words and 440MB of text. In addition to the blog posts, we obtained the comments posted in response to the posts—a total of 1.5M comments, 35M words, and 320MB of text.

Ad Corpus. We acquired a set of 18,500 advertisements which are currently used for the blogs in our collection (and for other web pages: the company

Title	ArtOlive - More than 2,250 Dutch Artists
Description	The platform for promoting, lending and selling contemporary art all over the Netherlands. Click to view the current collection of more than 2,250 artists, or read about buying and renting art.
URL	`www.galerie.nl`
Trigger Words	painting, sculpture, galleries, artist, artwork, studio, artists, studios, gallery
Title	Start dating on Lexa.nl
Description	It's time for a new start. About 30,000 profiles every month. Register now for free.
URL	`www.lexa.nl`
Trigger Words	dating, meeting, dreamgirl, contacts

Fig. 2. Sample advertisements from our collection.

that operates web-log.nl, Ilse Media BV, also hosts the largest Dutch search engine and a popular portal). In total, 1,650 different web sites are advertised in the collection, and 10,400 different "triggering words" are used. Figure 2 shows examples of the advertisements in our collection.

As Dutch is a compound-rich language, we used a compound-splitting technique that has shown substantial improvements in retrieval effectiveness compared to unmodified text [24] for all components of our method employing retrieval.

4.2 Experiments

Three methods were used to match ads to blog contents. As a baseline, we indexed all ads and used the blog post as a query, ranking the ads by their retrieval score; in addition, the appearance of a trigger word in the post was required. This is similar to the AAK ("match Ads And Keywords") method described in [9], except that we use the language modeling approach to information retrieval described in [20] for ranking the ads rather than a vector space one. This most likely improves the scores of the baseline, as the language modeling retrieval method we use has shown to achieve same-or-better scores compared to top-performing retrieval algorithms, and certainly outperforms the simpler vector space model [20]. We refer to this method as AAK.

To address the first of our main research questions (How effective are state-of-the-art ad placement methods on blogs?), we implemented the impedance coupling method AAK_EXP described in [9] (the acronym stands for "match Ad And Keywords to the EXPanded page"); this represents current state-of-the-art of content-based ad matching.[2] Again, we used the language modeling framework for the retrieval component in this method, which is likely to improve its performance.

Finally, we used the language modeling mixture method for ad placement described in Section 3. Since we did not have training material we could not tune the prior weights of the models, and used a naive weighting scheme as

---

[2] The authors of [9] implement a number of methods for ad matching; AAK_EXP and AAK_EXP_H are the top-performing methods, where AAK_EXP_H shows a minor advantage over AAK_EXP but requires an additional crawling step which we did not implement.

Method	Precision@1	Precision@3
AAK [9] (baseline)	0.18	0.18
AAK_EXP [9]	0.25 (+39%)	0.24 (+33%)
LANG_MODEL_MIX	0.28 (+55%)	0.29 (+61%)

Table 2. Ad-matching evaluation.

detailed in Section 3. Posterior weights were applied as described in Section 3, according to the coherency of the resulting models. We refer to this method as LANG_MODEL_MIX.

Assessment. For evaluation purposes we randomly selected a set of 103 blog posts as a test set. The top 3 advertisements selected by all three methods for each of these posts were assessed for relevance by two independent assessors. The assessors viewed the blog posts in their original HTML form (i.e., complete with images, links, stylesheets and other components); at the bottom of the page a number of advertisements were displayed in random order, where the method used to select an ad was not shown. The assessors were asked to mark an advertisement as "relevant" if it is likely to be of interest to readers of this blog, be they incidental visitors to the page or regular readers.

The level of agreement between the assessors was $\kappa = 0.54$. Due to this relatively low value, we decided to mark an advertisement "relevant" for a blog post only if both assessors marked it as relevant.[3]

4.3 Results

To evaluate the ad selection methods, we measured the precision levels for the top-ranked ad selected by the method, as well as the 3 top-ranked ads (a larger number of ads is likely to disturb visitors to the blog). Table 2 shows the average precision scores for all methods. All differences are strongly statistically significant using the sign test, with p values well below 0.001.

As shown in [9], the usage of the sophisticated query expansion mechanism of AAK_EXP yields a substantial improvement over the baseline. However, the improvement is somewhat lower than that gained for generic web pages: while the average improvement reported in [9] is 44%, in the case of blogs the average improvement is 36%. Usage of the LANG_MODEL_MIX method shows yet another substantial improvement, of the same order of magnitude, suggesting that this is a beneficial scheme for capturing a profile of the blog post for commercial purposes. Note that an improvement of $X\%$ in ad-matching can lead to an improvement of $X\%$ in the end result (in this case, sales from advertisements),

---

[3] The requirement that two independent assessors agree on an ad's relevance leads to more robust evaluation, but also reduces the scores, as fewer advertisements (on average) are marked as relevant. A different policy, marking an advertisement as relevant if *any* of the assessors decided it is relevant, boosts all scores by about 40%, but makes them less reliable.

unlike many other computational linguistic tasks where the effect of performance enhancements on the end result is not linear [11].

An in-depth analysis of the contribution of the different models to the outcome, as well as error analysis, is out of the scope of this paper, and will be made public separately.

5 Conclusions

Our aim in this work was two-fold: to determine the effectiveness of state-of-the-art ad placement methods on blogs (as opposed to general non-blog web pages), and to propose a blog-specific ad placement algorithm that builds on the intuition that a blog represents a person, not a single topic. We used manual assessments of a relatively large test set to compare our blog-specific method to a top performing state-of-the-art one—AAK_EXP. While AAK_EXP performs well, the richness of information in blogs enables us to significantly improve over it.

The success of our method is based on the use of properties which are relatively unique to blogs—the presence of a community, comments, the fact that the post itself is part of a blog, and so on. We believe that further improvements may be achieved by using non-blog specific features; among these are linguistic cues such as sentiment analysis (shown to improve other commercial-oriented tasks dealing with blogs [25]), as well as non-linguistic ones such as ad expansion, e.g., from the page pointed to by the ad [9]. Another interesting line of further work concerns the integration of additional knowledge about the blog reader, as mined from her clickstream or her own blog.

References

1. Herring, S., Scheidt, L., Bonus, S., Wright, E.: Bridging the gap: A genre analysis of weblogs. In: HICSS. (2004)
2. TREC: Blog track (2006) URL: `http://trec.nist.gov`.
3. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: Proceedings KDD '05, New York, NY, USA, ACM Press (2005) 78–87
4. Mishne, G., de Rijke, M.: Deriving wishlists from blogs: Show us your blog, and we'll tell you what books to buy. In: Proceedings WWW2006. (2006)
5. Lingamneni, S.: Predicting the future of internet advertising (2004) `http://www.stanford.edu/group/boothe/0405/PWR-Lingamneni.pdf`.
6. Wang, C., Zhang, P., Choi, R., Daeredita, M.: Understanding consumers attitude toward advertising. In: Eighth Americas Conference on Information Systems. (2002) 1143–1148

7. Bhargava, H.K., Feng, J.: Paid placement strategies for internet search engines. In: Proceedings of the eleventh international conference on World Wide Web. (2002) 117–123

8. Novak, T.P., Hoffman, D.L.: New metrics for new media: toward the development of web measurement standards. World Wide Web J. **2** (1997) 213–246

9. Ribeiro-Neto, B., Cristo, M., Golgher, P., de Moura, E.S.: Impedance coupling in content-targeted advertising. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (2005) 496–503

10. Langheinrich, M., Nakamura, A., Abe, N., Kamba, T., Koseki, Y.: Unintrusive customization techniques for web advertising. Comput. Networks **31** (1999) 1259–1272

11. Tau-Wih, W., Goodman, J., Carvalho, V.: Finding advertising keywords on web pages. In: Proceedings of the World Wide Web Conference 2006, Edinburgh, Scotland (2006)

12. Rosenfeld, R.: Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE **88** (2000)

13. Ponte, J., Croft, W.: A language modeling approach to information retrieval. In: Proceedings SIGIR '98, New York, NY, USA, ACM Press (1998) 275–281

14. Brown, R., Frederking, R.: Applying statistical english language modeling to symbolic machine translation. In: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95). (1995) 221–239

15. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19** (1993) 61–74

16. Winer, D.: What makes a weblog a weblog? (2003) `blogs.law.harvard.edu/whatMakesAWeblogAWeblog`, accessed November 2005.

17. Gumbrecht, M.: Blogs as "protected space". In: WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004. (2004)

18. Golder, S., Huberman, B.: The structure of collaborative tagging systems. Journal of Information Science (2006)

19. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. In: Proceedings WWW2003, New York, NY, USA, ACM Press (2003) 568–576

20. Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis, Enschede (2001)

21. Lavrenko, V.: Optimal Mixture Models in IR. In: ECIR2002. (2002) 193–212

22. Kalai, A., Chen, S., Blum, A., Rosenfeld, R.: On-line algorithms for combining language models. In: Proceedings of the International Conference on Accoustics, Speech, and Signal Processing (ICASSP '99). (1999)

23. Turney, P.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: ECML. (2001) 491–502

24. Hollink, V., Kamps, J., Monz, C., de Rijke, M.: Monolingual document retrieval for European languages. Information Retrieval **7** (2004) 33–52

25. Mishne, G., Glance, N.: Predicting movie sales from blogger sentiment. In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006). (2006)