The Happy Searcher: Challenges in Web Information Retrieval

Mehran Sahami Vibhu Mittal Shumeet Baluja Henry Rowley

Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
{sahami, vibhu, shumeet, har}@google.com

Abstract. Search has arguably become the dominant paradigm for finding information on the World Wide Web. In order to build a successful search engine, there are a number of challenges that arise where techniques from artificial intelligence can be used to have a significant impact. In this paper, we explore a number of problems related to finding information on the web and discuss approaches that have been employed in various research programs, including some of those at Google. Specifically, we examine issues of such as web graph analysis, statistical methods for inferring meaning in text, and the retrieval and analysis of newsgroup postings, images, and sounds. We show that leveraging the vast amounts of data on web, it is possible to successfully address problems in innovative ways that vastly improve on standard, but often data impoverished, methods. We also present a number of open research problems to help spur further research in these areas.

1 Introduction

Search engines are critically important to help users find relevant information on the World Wide Web. In order to best serve the needs of users, a search engine must find and filter the most relevant information matching a user's query, and then present that information in a manner that makes the information most readily palatable to the user. Moreover, the task of information retrieval and presentation must be done in a scalable fashion to serve the hundreds of millions of user queries that are issued every day to a popular web search engines such as Google.

In addressing the problem of information retrieval on the web, there are a number of challenges in which Artificial Intelligence (AI) techniques can be successfully brought to bear. We outline some of these challenges in this paper and identify additional problems that may motivate future work in the AI research community. We also describe some work in these areas that has been conducted at Google.

We begin by briefly outlining some of the issues that arise in web information retrieval that showcase its differences with research traditionally done in Information Retrieval (IR), and then focus on more specific problems. Section 2 describes the unique properties of information retrieval on the web. Section 3 presents a statistical method for determining similarity in text motivated by both AI and IR methodologies.

Section 4 deals with the retrieval of UseNet (newsgroups) postings, while Section 5 addresses the retrieval of non-textual objects such as images and sounds. Section 6 gives a brief overview of innovative applications that harness the vast amount of text available on the Web. Finally, Section 7 provides some concluding thoughts.

2 Information Retrieval on the Web

A critical goal of successful information retrieval on the web is to identify which pages are of high quality and relevance to a user's query. There are many aspects of *web* IR that differentiate it and make it somewhat more challenging than traditional problems exemplified by the TREC competition. Foremost, pages on the web contain links to other pages and by analyzing this web graph structure it is possible to determine a more global notion of page quality. Notable early successes in this area include the PageRank algorithm [1], which globally analyzes the entire web graph and provided the original basis for ranking in the Google search engine, and Kleinberg's HITS algorithm [2], which analyzes a local neighborhood of the web graph containing an initial set of web pages matching the user's query. Since that time, several other linked-based methods for ranking web pages have been proposed including variants of both PageRank and HITS [3][4], and this remains an active research area in which there is still much fertile research ground to be explored.

Besides just looking at the link structure in web pages, it is also possible to exploit the anchor text contained in links as an indication of the content of the web page being pointed to. Especially since anchor text tends to be short, it often gives a concise human generated description of the content of a web page. By harnessing anchor text, it is possible to have index terms for a web page even if the page contains only images (which is seen, for example, on visually impressive home pages that contain no actual text). Determining which terms from anchors and surrounding text should be used in indexing a page presents other interesting research venues.

2.1 Adversarial Classification: Dealing with Spam on the Web

One particularly intriguing problem in web IR arises from the attempt by some commercial interests to unduly heighten the ranking of their web pages by engaging in various forms of *spamming* [5]. One common method of spamming involves placing additional keywords (or even entire dictionaries) in invisible text on a web page so that the page potentially matches many more user queries, even if the page is really irrelevant to these queries. Such methods can be effective against traditional IR ranking schemes that do not make use of link structure, but have more limited utility in the context of global link analysis. Realizing this, spammers now also utilize *link spam* where they will create large numbers of web pages that contain links to other pages whose rankings they wish to raise.

Identifying such spam in both text-based and linked-based analyses of the web are open problems where AI techniques such as Natural Language Processing (NLP) and Machine Learning (ML) can have a direct impact. For example, statistical NLP

methods can be used to determine the likelihood that text on a web page represents "natural" writing. Similarly, classification methods can be applied to the problem of identifying "spam" versus "non-spam" pages, where both textual and non-textual (e.g., link) information can be used by the classifier.

Especially interesting is that such classification schemes must work in an *adversarial* context as spammers will continually seek ways of thwarting automatic filters. Adversarial classification is an area in which precious little work has been done, but effective methods can provide large gains both for web search as well as other adversarial text classification tasks such as spam filtering in email [6].

2.2 Evaluating Search Results

Even when advances are made in the ranking of search results, proper evaluation of these improvements is a non-trivial task. In contrast to traditional IR evaluation methods using manually classified corpora such as the TREC collections, evaluating the efficacy of web search engines remains an open problem and has been the subject of various workshops [7][8]. Recent efforts in this area have examined interleaving the results of two different ranking schemes and using statistical tests based on the results users clicked on to determine which ranking scheme is "better" [9]. There has also been work along the lines of using decision theoretic analysis (i.e., maximizing users' utility when searching, considering the relevance of the results found as well as the time taken to find those results) as a means for determining the "goodness" of a ranking scheme. Commercial search engines often make use of various manual and statistical evaluation criteria in evaluating their ranking functions. Still, principled automated means for large-scale evaluation of ranking results are wanting, and their development would help improve commercial search engines and create better methodologies to evaluate IR research in broader contexts.

3 Using the Web to Create "Kernels" of Meaning

Another challenge in web search is determining the relatedness of fragments of text, even when the fragments may contain few or no terms in common. In our experience, English web queries are on average two to three terms long. Thus, a simple measure of similarity, such as computing the cosine of the terms in both queries, is very coarse and likely to lead to many zero values. For example, consider the fragments "Captain Kirk" and "Star Trek". Clearly, these two fragments are more semantically similar than "Captain Kirk" and "Fried Chicken", but a simple term-based cosine score would give the same (zero) value in both cases.

Generalizing this problem, we can define a real-valued kernel function K(x, y), where x and y are arbitrary text fragments. Importantly, we note that K can utilize external resources, such as a search engine in order, to determine a similarity score[1]. To this end, we can perform query expansion [10] on both x and y using the results of

[1] We could define K(x, y, S) where S represents the search engine used. However, since S generally remains constant, we can define K with respect to just the parameters x and y.

a search engine and then compute the cosine between these expanded queries. More formally, let QE(t) denote the query expansion of text t, where (for example) we could define QE(t) as the centroid of the TFIDF vector representations of the top 30 documents returned by a search engine in response to query t. We can now define K(x, y) as the cosine between QE(x) and QE(y). Illustratively, we obtain the following results with such a kernel function, anecdotally showing its efficacy:

K("Captain Kirk", "Mister Spock") = 0.49
K("Captain Kirk", "Star Trek") = 0.38
K("Captain Kirk", "Fried Chicken") = 0.02

While such a *web contextual kernel* function has obvious utility in determining the semantic relatedness of two text fragments by harnessing the vast quantities of text on the web, open research issues remain. For example, future research could help identify more effective text expansion algorithms that are particularly well suited to certain tasks. Also, various methods such as statistical dispersion measures or clustering could be used to identify poor expansions and cases where a text fragment may have an expansion that encompasses multiple meanings (e.g., an expansion of "Michael Jordan" including terms both about the researcher and the basketball star).

4 Retrieval of UseNet Articles

One of the less visible document collections in the context of general purpose search engines is the UseNet archive, which is conservatively estimated to be at least 800 million documents. The UseNet archive, mostly ignored in traditional academic IR work—with the one exception of the 20 newsgroups data set used in text classification tasks—is extremely interesting. UseNet started as a loosely structured collection of groups that people could post to. Over the years, it evolved into a large hierarchy of over 50,000 groups with topics ranging from sex to theological musings.

IR in the context of UseNet articles raises some very interesting issues. As in the case of the Web, spam is a constant problem. However, unlike the web, there is no clear concept of a home page in UseNet. For example, what should the canonical page for queries such as "IBM" or "Digital Cameras" be? One previously explored possibility is to address retrieval in UseNet as a two stage IR problem: (1) find the most relevant newsgroup, and (2) find the most relevant document within that newsgroup. While this may appear to be a simple scheme, consider the fact that there are at least 20 newsgroups that contain the token "IBM". This leads us to the problem of determining whether the canonical newsgroup should be based on having "IBM" at the highest level (i.e., `comp.ibm.pc`), the group with the most subgroups underneath it (i.e., `comp.sys.ibm.*`), or simply the most trafficked group. Still, other questions arise, such as whether moderated newsgroups should given more weight that un-moderated newsgroups or if the *Big-8* portion of the UseNet hierarchy should be considered more credible than other portions.

At the article or posting level, one can similarly rank not just by content relevance, but also take into account aspects of articles that not normally associated with web pages, such as temporal information (when a posting was made), thread information,

the author of the article, whether the article quotes another post, whether the proportion of quoted content is much more than the proportion of original content, etc. Moreover, recognizing that certain postings may be FAQs or "flames" would also aid in determining the appropriate ranking for an article. Along these lines, previous research has examined building models of newsgroups, communication patterns within message threads, and language models that are indicative of content [11][12][13]. Still, questions remain of how to go about using such factors to build an effective ranking function and how to display these results effectively to users.

Furthermore, one can also attempt to compute the inherent quality or credibility level of an author independent of the query, much as PageRank [1] does for the Web. Such a computation would operate on a graph of relatively modest size since, for example, if we were to filter authors to only those that had posted at least twice in a year to the same newsgroup, we would be left with only on the order of 100,000 authors. This is a much more manageable size than the web graph which has several billion nodes. Computing community structures—rather than pure linear structures as in posting threads—can also generate interesting insights as to how various authors and groups participate in and influence discussions.

One of the most comprehensive studies on bulletin board postings (similar to UseNet) is the Netscan project [11]. This work examined characteristics of authors and posting patterns, such as identifying characteristics of people who start discussions, people who "flame", people who cross-post to multiple newsgroups, people who spam, people who seem to terminate threads, etc. More recently, work on filtering technologies in the context of information retrieval [14] has also focused attention on building better models of the likely content in messages and routing them to appropriate people, bringing together work on user modeling, IR, and text analysis.

An advantage of working with the UseNet archive is the fact that it alleviates many of the infrastructural problems that might otherwise slow research in the web domain, such as building HTML parsers, properly handling different languages and character sets, and managing the exceptional volume of available data (even small potions of the Web would require several hundred gigabytes to store). Contrastingly, much of the older UseNet posting archive was previously available on a few CD-ROMs, making the archive relatively easy to store, index and process on a single machine. More recently, researchers have started looking at an even smaller scale problem: culling information from bulletin board postings and trying to ascribe a quality level to the information contained therein. For example, Arnt and Zilberstein [13] analyzed postings on the Slashdot bulletin board (a discussion forum predominated by technology savvy readers), attempting to learn the moderation system used. Slashdot moderators assign both a genre label— such as "informative", "funny", etc.—and a score between -1 and +5 indicating their view on how relevant a posting is. Given these score and label pairs, it is a challenging task to use the rich structure of the domain (i.e., author information, posting content, thread history, etc.) to predict both the label and score for new postings. More generally, improving ranking methods for UseNet or bulletin board postings is an open area of research with many interesting similarities to the web, but also with very many significant differences that make it a fascinating subject of further study.

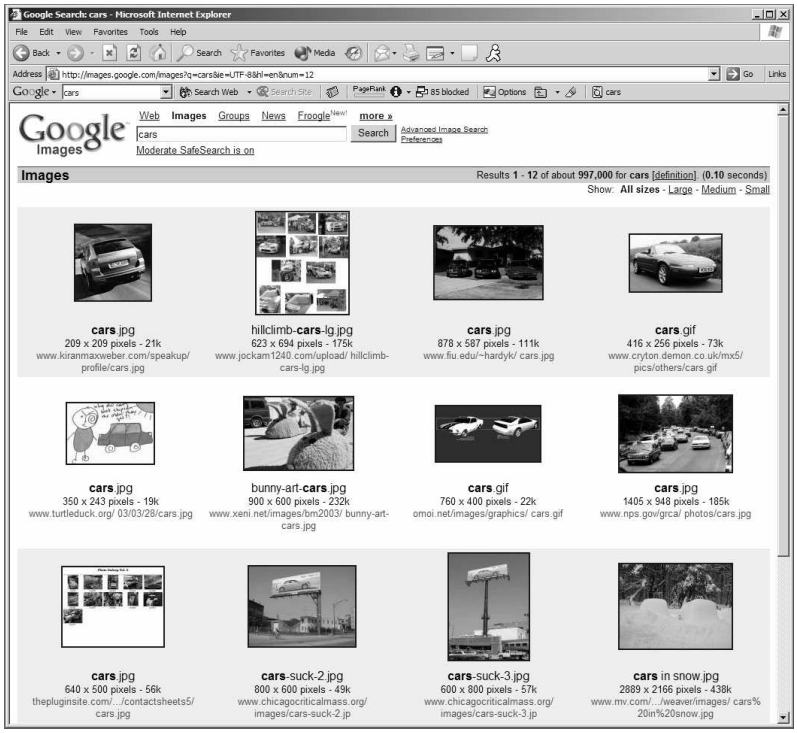

Fig. 1: 12 Results obtained by searching Google-Images for "Cars"

5 Retrieval of Images and Sounds

With the proliferation of digital still and video cameras, camera phones, audio recording devices, and mp3 music, there is a rapidly increasing number of non-textual "documents" available to users. One of the challenges faced in the quest to organize and make useful all of the world's information, is the process by which the contents of these non-textual objects should indexed. An equally important line of study (although not a focus of this paper) is how to present the user with intuitive methods by which to query and access this information.

The difficulties in addressing the problem of non-textual object retrieval are best illustrated through an example. Figure 1 shows 12 results obtained by searching Google's image repository for "cars". Note the diverse set of content related to cars that is present. In the first 12 results, we see everything from different car poses, pictures of cars on billboards, cars barely visible through the snow, cars for parades, and even hand drawn illustrations. In addressing this sort of diversity, we presently give three basic approaches to the task of retrieving images and music.

1. ***Content Detection:*** For images, this method means that the individual objects in the image are detected, possibly segmented, and recognized. The image is then labeled with detected objects. For music, this method may include recognizing the instruments that are played as well as the words that are said/sung, and even determining the artists. Of the three approaches, this is the one that is the furthest from being adequately realized, and involves the most signal processing.

2. ***Content Similarity Assessment:*** In this approach, we do not attempt to recognize the content of the images (or audio clips). Instead, we attempt to find images (audio tracks) that are similar to the query items. For example, the user may provide an image (audio snippet) of what the types of results that they are interested in finding, and based on low-level similarity measures, such as (spatial) color histograms, audio frequency histograms, etc, similar objects are returned. Systems such as these have often been used to find images of sunsets, blue skies, etc. [15] and have also been applied to the task of finding similar music genres [16].

3. ***Using Surrounding Textual Information:*** A common method of assigning labels to non-textual objects is to use information that surrounds these objects in the documents that they are found. For example, when images are found in web documents, there is a wealth of information that can be used as evidence of the image contents. For example, the site on which the image appears (for example an adult site or a site about music groups, TV shows, etc.), how the image is referred to, the image's filename, and even the surrounding text all provide potentially relevant information about the image.

All of these approaches can, of course, be used in conjunction with each other, and each provides a fairly diverse set of benefits and drawbacks. For example, surrounding textual information is the easiest method to use; however it is the most susceptible to misclassification of the image content, due to both errors and malicious web site designers. Content Similarity Assessment can provide some indication of the image content, but is rarely able in practice to find particular objects or particular people. Content Detection is the only method that attempts to recognize the objects in the scene; however, building detectors for arbitrary objects is a time consuming task that usually involves quite a bit of custom research for each object. For example, the most studied object detection domain to date is finding faces in images, and work has continued on improving the quality for almost a decade [17][18][19][20]. Work in using these systems to detect people (beyond just finding faces) and cars is progressing [21][22]; extending to arbitrary objects is also the focus of a significant amount of research.

Beyond assigning labels to images, there are a variety of other topics that must be addressed in deciding which images to present to the user. For example, should multiple copies of the same image be presented? What about near-duplicates? Eliminating near-duplicates involves not only comparing the images to find identical copies, but also developing automatic methods to ignore insignificant variations – such as those due to compression formats, scanner calibration error, and small corruptions in files. Another topic that must be addressed is what order to present the

images. Is there one ordering that is better than another? Perhaps the relevance of the page on which the images are found should play a factor in the order assessment. Finally, looking into the future, how many of these ideas can be extended to video retrieval? Combining the audio track from videos with the images that are being displayed may not only provide additional sources of information on how to index the video, but also provide a tremendous amount of (noisy) training data for training object recognition algorithms en masse.

6 Harnessing Vast Quantities of Data

Even with the variety of research topics discussed previously, we are only still scratching the surface of the myriad of issues that AI technologies can address with respect to web search. One of the most interesting aspects of working with web data is the insight and appreciation that one can get for large data sets. This has been exemplified by Banko and Brill in the case of word sense disambiguation [23], but as a practical example, we also briefly discuss our own experiences in two different contexts at Google: Spelling Correction and Query Classification.

Spelling Correction. In contrast to traditional approaches which solely make use of standard term lexicons to make spelling corrections, the Google spelling corrector takes a Machine Learning approach that leverages an enormous volume of text to build a very fine grained probabilistic context sensitive model for spelling correction. This allows the system to recognize far more terms than a standard spelling correction system, especially proper names which commonly appear in web queries but not in standard lexicons. For example, many standard spelling systems would suggest the text "Mehran Sahami" be corrected to "Tehran Salami", being completely ignorant of the proper name and simply suggesting common terms with small edit distance to the original text. Contrastingly, the Google spelling corrector does not attempt to correct the text "Mehran Sahami" since this term combination is recognized by its highly granular model. More interesting, however, is the fact that by employing a *context sensitive* model, the system will correct the text "Mehran *Salhami*" to "Mehran *Sahami*" even though "Salami" is a common English word and is the same edit distance from "Salhami" as "Sahami." Such fine grained context sensitivity can only be achieved through analyzing very large quantities of text.

Query Classification into the Open Directory Project. The Open Directory Project (ODP) (`http://dmoz.org/`) is a large open source topic hierarchy into which web pages have been manually classified. The hierarchy contains roughly 500,000 classes/topics. Since this is a useful source of hand-classified information, we sought to build a query classifier that would identify and suggest categories in the ODP that would be relevant to a user query. At first blush, this would appear to be a standard text classification task. It becomes more challenging when we consider that the "documents" to be classified are user queries, which have an average length of just over two words. Moreover, the set of classes from the ODP is much larger than any previously studied classification task, and the classes are non-mutually exclusive

which can create additional confusion between topics. Despite these challenges, we have available roughly four million pre-classified documents, giving us quite a substantial training set.

We tried a variety of different approaches that explored many different aspects of the classifier model space: independence assumptions between words, modeling word order and dependencies for two and three word queries, generative and discriminative models, boosting, and others. The complete list of methods compared is not included since some portions of the study were conducted in an iterative piecemeal fashion, so a direct comparison of all methods applied to all the data is not possible to provide. Nevertheless, we found that the various algorithms performed as expected relative to previously published results in text classification when training data set sizes were small. Interestingly, as we steadily grew the amount of data available for training, however, we reached a critical point at which most of the algorithms were generally indistinguishable in performance. Even more interesting was the fact that as we moved substantially beyond this critical point by adding even more training data, Naïve Bayes (with a few very minor modifications to take into account the confidence associated with the classification and the use of a separate model for single word queries), outperformed—by several percentage points in accuracy—every other algorithm employed, even after substantial effort was placed into making them better. Furthermore, most probability smoothing techniques, which generally seem to help in limited data situations, either showed no appreciably improvements or actually decreased performance in the data rich case for Naïve Bayes.

While the set of alternative algorithms used was by no means exhaustive, and the results here are still somewhat anecdotal, we hypothesize that, as in the case of the Banko and Brill study, an abundance of data often can, and usually does, make up for weaker modeling techniques. This perspective can be unusually liberating—it implies that given enough training data, the simpler, more obvious solutions can work, perhaps even better than more complex models that attempt to compensate for lack of sufficient data points.

7 Conclusions

Web information retrieval presents a wonderfully rich and varied set of problems where AI techniques can make critical advances. In this paper, we have presented a number of challenges, giving an (admittedly brief) overview of some approaches taken toward these problems and outlining many directions for future work. As a result, we hope to stimulate still more research in this area that will make use of the vast amount of information on the web in order to better achieve the goal of organizing the world's information and making it universally accessible and useful.

References

1. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proc. of the 7th International World Wide Web Conference (1998) 107-117

2. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM **46**(5) (1999) 604-632
3. Bharat, K., Henzinger, M.R.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proc. of the 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval (1998) 104-111
4. Tomlin, J.A.: A New Paradigm for Ranking Pages on the World Wide Web. In: Proc. of the 12th International World Wide Web Conference (2003) 350-355
5. Henzinger, M.R., Motwani, R., Silverstein, C.: Challenges in Web Search Engines. In: Proc. of the 18th International Joint Conference on Artificial Intelligence (2003) 1573-1579
6. Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E.: A Bayesian Approach to Filtering Junk E-Mail. In: Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05 (1998)
7. Dumais, S., Bharat, K., Joachims, T., Weigend, A. (eds.): Workshop on Implicit Measures of User Interests and Preferences at SIGIR-2003 (2003).
8. Agosti, M., and Melucci, M. (eds.): Workshop on Evaluation of Web Document Retrieval at SIGIR-1999 (1999)
9. Joachims, T.: Evaluating Retrieval Performance Using Clickthrough Data. In Proc. of the SIGIR-2002 Workshop on Mathematical/Formal Methods in Information Retrieval (2002)
10. Mitra. M., Singhal, A., and Buckley, C.: Improving Automatic Query Expansion. In: Proc. of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (1998) 206-214
11. Smith, M., Kollock, P.: Communities in Cyberspace: Perspectives on New Forms of Social Organization. Routledge Press, London (1999)
12. Fiore, A., Tiernan, S.L., Smith, M.: Observed Behavior and Perceived Value of Authors in Usenet Newsgroups: Bridging the Gap, In: Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems (2002) 323-330
13. Arnt, A., and Zilberstein, S.: Learning to Perform Moderation in Online Forums. In: Proc. of the IEEE/WIC International Conference on Web Intelligence (2003)
14. Zhang, Y., Callan, J., Minka, T.P.: Novelty and Redundancy Detection in Adaptive Filtering. In: Proc. of the 25th International ACM-SIGIR Conference on Research and Development in Information Retrieval (2002)
15. Smith, J.R., Chang, S.F.: Tools and Techniques for Color Image Retrieval. In: Proc. of SPIE Storage and Retrieval for Image and Video Databases, Vol. 2670. (1996) 426-437
16. Berenzweig, A., Logan, B., Ellis, D., Whitman, B.: A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. In: Proc. of the 4th International Symposium on Music Information Retrieval (2003)
17. Wu, J., Rehg, J.M., Mullin, M.D.: Learning a Rare Event Detection Cascade by Direct Feature Selection. In: Advances in Neural Information Processing Systems 16 (2004)
18. Sung, K., Poggio, T.: Learning Human Face Detection in Cluttered Scenes. In Proc. of Intl. Conf. on Computer Analysis of Image and Patterns (1995)
19. Rowley, H.A., Baluja, S., Kanade, T.: Neural Network-based Face Detection. IEEE Trans. On Pattern Analysis and Machine Intelligence **20**(1) (1998) 23-38
20. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2001) 511-518
21. Schneiderman, H., Kanade, T.: A Statistical Model for 3D Object Detection Applied to Faces and Cars. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (2000)
22. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. Mitsubishi Electric Research Lab Technical Report. TR-2003-90 (2003)
23. Banko, M., Brill, E.: Mitigating the Paucity of Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for NLP. In: Proc. of the Conference on Human Language Technology (2001)