Autovita: a 360 degree view of Person Information on the WWW

Abstract

Search for a person over the WWW, is usually targeted to find personal information or to get a comprehensive overview of his/her web presence. Since, current day search engines still treat any search query as keywords, they fail to provide a comprehensive and quick summary for person related searches, instead returning a huge list of keyword based matching documents. It is largely left to the user to decide which page he is interested in and also comprehend and analyze the bits and pieces of information redundantly and sparsely distributed across various documents. In this paper, we discuss our system Autovita which mines person information from the web to provide a comprehensive and non-redundant summary for different dimensions of interest like background information, finance issues, colleague information etc. Such a 360 degree view for a person related search helps in quick identification and analysis of a person's web presence.

1 Introduction

The Web is a vast repository of information and data that grows continuously. Information traditionally published in other media (e.g. manuals, brochures, magazines, books, newspapers, etc.) is now increasingly published and is distributed through the web. Today, the web contains more than one a few billion pages. The sheer size and explosive growth of the web has created the need for tools and methods that can automatically search, index, access, extract and recombine information and knowledge that is publicly available from web resources.

To deal with this excessive information overload, we have search engines, that spider a large portion of the web and provide search results in sparklingly less time. Search engines have become a part of our daily life. We extensively use search engines to find information related to people. A search for a person over the web can broadly be classified as having two intentions - one for seeking the personal information like address, profile, job details or two, to know more about the person in various dimensions like relation with other people, institutions, organizations or participation in events and any finance related issues etc. In order to effectively address the information need of users when querying for person related searches, it would be useful to have a comprehensive 360 degree overview of the person, with non-redundant information in these various dimensions.

The extraction of personal information from personal pages is generally quite easy. For example a generic IE system easily spots email addresses and telephone numbers, etc. and they tend to be unique in the page. For other information (e.g. the position of a person, such as professor, researcher, etc.) it is possible to use wrappers trained to extract information from homepages (Kushmerick et al., 1997). There are systems specializing in homepage identification and extraction of information from homepages [1].

Extraction of information to satisfy the second kind of information need like relation to events, press-releases, finances, projects, publicity, association with organizations, relation with other people etc is difficult, firstly as they are not always published on a person's homepage and secondly as it is still linguistically challenging to do so. This information however is present in the word wide web, along with plenty of other possible information in superficial formats as redundant copies and poses a challenge for information extraction. Redundancy is given by the presence of multiple cita-

[1] http://www.zoominfo.com/

tions of the same facts in different superficial formats and has been used for several tasks such as improving question answering systems (Dumais et al., 2002) and performing information extraction using machine learning (Mitchell, 2001).

Current day search engines do a good job in retrieving all the web documents related to the person name provided based on a keyword search. These documents may sometimes to a large extent, contain web pages related to other like-named people or namesakes that are different from the one in the query. It is largely left to the user to decide which page he is interested in and filter the information accordingly. Even though, a web user does take the effort of identifying his required person in context, he still will be left with a huge list of documents, with bits and pieces of information redundantly and sparsely distributed across various documents. There is yet another problem of mis-information being published intentionally or unintentionally across these various documents. It would be interesting and useful to have a wholesome view of a person at one single place, and also have some credibility associated with such information for quick cross verification either by the end-user or the actual owner of the information.

In this paper, we discuss our Autovita system which mines person information from the web to provide comprehensive 360 degree 'view' for person related searches. We use a real world search engine results for a person's name and perform simple text mining and information extraction techniques to build a comprehensive view of the person. We propose a technique that uses the homepage of a person for disambiguating the webpages of a person from among an enormous list of results returned by a search engine for a person name. We also propose and use an adapted similarity metric that makes use of named entity tags to effectively extract a set of non-redundant sentences related to a person that constitute a 'view' for the person. A 'view' for a person is any of the dimensions of interest in 'person related searches' like - background information, events participated, organizational relations , colleagues, financial appearances on the WWW etc. Finally, we use the redundant appearances on the WWW to predict outliers in the facts identified.

2 Related Work

Mining websites to collect information has been explored in areas of Text Mining, Information retrieval and extraction(Grishman, 1997) and Natural language processing. These areas have been extensively explored in the past few decades and a lot of the techniques proposed and applied in the context of WWW (Eikvil, 1999). Our work builds on well know techniques for text summarization (Mani and (editors), 1999). However, there are important practical differences between the traditional task of summarizing a document, and our problem of summarizing Web pages of a person. In particular, traditional summarization is not progressive. A document is summarized, and the user decides whether to read the full document. Web pages of a person have very diverse content , it does not make sense to summarize the entire set of pages for the person as one unit. Instead we take the approach of identifying the dimensions in which we generate our summary. Systems like OCELOT (A.L. Berger, 2000) and MEAD (Radev et al., 2004) have attempted multidocument summarization techniques for the WWW, but the domain that we are working in - person related searches, has the additional task of identifying and pruning namesakes when searched for a person name on the WWW.

Autovita is closely related to two projects Armadillo (Fabio Ciravegna and Wilks, 2004) and KnowItAll (Etzioni et al., 2004). Armadillo project uses adaptive techniques for information extraction (Dingli et al., 2003) and integrates information from large respositories, usually websites. Our approach is motivated by the Armadillo project, but is targeted for person related searches on the WWW, which usually return a lot of results that may not belong to the user, which needs to be identified. The KnowItAll project (Etzioni et al., 2004), aims to automate the tedious process of extracting large collections of facts from the web in an autonomous, domain-independent, and scalable manner. Although, we borrow the spirit of pattern based knowledge extraction from the KnowItAll project, both these projects are widely different because of the domains they work in. Where as KnowItAll intends to provide a complete view of knowledge over the WWW, our project intends to explore the possiblity of providing a comprehensive view for person related searches.

The work related to the sub parts of the Autovita

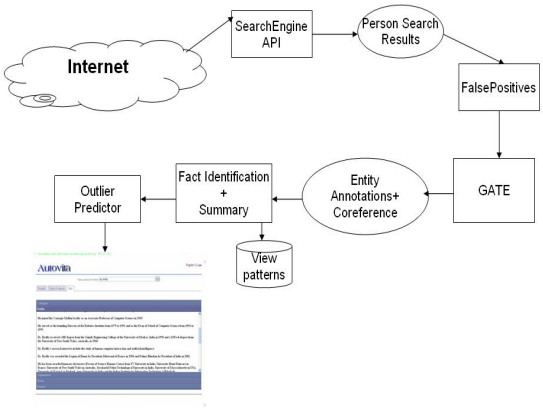

Figure 1: System architecture

system is discussed wherever introduced.

3 Autovita

The goal of Autovita, is to provide a comprehensive 360 degree view for all person related searches. A person first registers to the system and provides details that act as prior knowledge that can help for disambiguating the search results. We use this information to obtain further information related to the person from the WWW. Although, we require the user to provide his homepage for best possible extraction of information, our experiments have shown that even fewer details like 'organization name', colleagues information etc help in the processing of information for the person.

The architecture of the system is shown in Figure 1. The architecture of Autovita is modular and amenable to modification. We follow a pipeline and filter architecture as it is quite suitable for our system with a lot of sequential batch processing taking place. Each filter in the system processes the data and either adds more information or prunes existing information, finally creating a comprehensive and non-redundant view of information when searched for a related person. The searches and intermediate processing results are cached in databases and filesystems to save future processing times.

In the remainder of the section, we discuss each of the sub systems and the data flow in detail.

3.1 Gathering information

The task of aggregating the complete data of a person from the WWW and formatting it as a summary seems daunting. Search engines come to the resuce here. Instead of performing a pull approach of crawling the entire WWW and extracting information, we have adopted a 'push' approach which only crawls for a particular person registered with our system. We first use search engines to extract results related to a person using his/her name. We use Google APIs and extract the top 1000 results for a person or top 10% of the results whichever is more. Search results returned usually contain "approximate replicas". We identify such duplicates before proceeding to the next stages primarily to save time on subsequent phases which are computationally expensive. We would like to ensure that the algorithms are run only on a non-redundant set of data. Upon downloading these results, we parse the HTML to extract content from the pages using a standard parser like JTidy [2].

3.2 False positives Identification

Most of the result pages returned by search engines for a particular person search, may not correspond to the person queried for and could possibly refer to another person having the same name called a 'namesake'. But how would the system identify whether certain Web pages are about the person in question or a different person with the same name?

For example, consider Raj Reddy, a University Professor at Carnegie Mellon University and a world reknowned researcher in Computer Science. When the query "Raj Reddy" is issued as a query to Google, most of the pages retrieved are actually related to him; however, there are also two students, a software consultant, a professor and a few others. If we are looking for information about a particular person, we want to filter out information about other namesakes, while also preserving the maximum amount of relevant information. It is sometimes quite difficult to determine if a page is about a particular person or not.

The problem of false positive identification is analogous to the problem of disambiguating web presence identification proposed in (Bekkerman and McCallum, 2005). We approach this problem by considering the information available on the home page of the person submitted at the

[2] http://jtidy.sourceforge.net/

time of registration with our system. Other namesakes(different person having same names) on the web are disambiguated using the information from the home page. Although this problem is clearly a classification problem ie one of that classifying any given page as belonging to the user or not, we could not use standard machine learning algorithms because the training data is very small (homepage).

Our approach is to first identify a set of key phrases from the background knowledge about the person. The neceessary background information is the home page of the person manually given to the system. More background information if given can be incorporated although not mandatary. Then, given any web page, we compute weights for each document using the key phrases extracted from the document.

In the rest of this subsection, we describe in detail our procedure of identfying false positives.

3.2.1 Identifying False Positives

Identifying false positives comprises three steps. 1. Feature Exraction 2. Document Weighting 3. Pruning and removing false positives

1. Feature Extraction: We first extract features from the documents. We consider all the nouns and verb phrases in the documents as the features of the document. We use ltchunker of Edinburgh University [3] for chunking the sentences in the downloaded result documents. The set of all the features f_h in the home page H comprise F_H.

2. Document Weighting: The weight W_{D_i} of a document D_i is computed as the sum of similarity between the features in the home page H and features in the document. For computing the similarity between the features, we considered a new similarity metric sim_{fp}. It is defined as shown in Eq (2). A value of 1 is assigned if the words are equal. Otherwise, simiarity using wordnet is calculated. Otherwise, a value of 0 is assigned. For computing the similarity using wordnet ($sim_{wn}(w_1, w_2)$), we measure the distance between the words in the wordnet heirarchy.

$$W_{D_i} = Sim_h(D_i, H) = \sum_{f_h \in F_H} \sum_{f_d \in D_i} sim_{fp}(f_h, f_d) \quad (1)$$

$$sim_{fp}(w_1, w_2) \begin{cases} 1 & \text{if } w_1 = w_2 \\ sim_{wn}(w_1, w_2) & \text{if found} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

3. Pruning and removing false positives: Pruning the false positives is done by selecting documents above a threshold. With a higher threshold, the recall reduces and precision increases. We have chosen a threshold of 0.2 based on experimentation.

3.3 Processing the Results

3.3.1 Named Entity Extraction

Developed at the University of Sheffield, ANNIE a Nearly New Information Extraction System exists inside the larger GATE infrastructure designed for developing and deploying software components that process human language (Cunningham et al., 2002). For the present project, we have used ANNIE as a named entity recognizer. Although ANNIE can provide many different types of annotations useful for many different applications, we rely upon ANNIE to provide annotations for people names, locations, dates, currency, organization and percentage .

ANNIE's information extraction though works well, is prone to errors. Reason for such errors is due to the fact that text on the WWW is very informal and proper names on the WWW belong to different communities and parts of the world and any amount of exhaustive listing of such names will not improve the tagging of proper names during Information extraction. Hence, we write wrappers that work around this problem to normalize the extracted entities, purely based on statistical counts and approximate expansion of the entity boundaries tagged by ANNIE.

We also make use of the orthographical coreference component, pronominal and nominal coreference resolution components built into the ANNIE in GATE framework to resolve coreferences.

3.4 Fact identification

The extraction of facts related to a person from his personal pages is generally considered as a straight forward information extraction task. For example a generic IE system easily spots email addresses and telephone numbers, etc. and they tend to be unique in the page. However, extraction of other pieces of information related to events,

[3] http://www.ltg.ed.ac.uk/software/chunk/index.html

press-releases, finances, association with organizations, relation with other people etc are difficult to find as they do not show up on a user's homepage or on its accompanying links.

In the remainder of the section we discuss how we extract such information and summarize it to provide a comprehensive view of person's information. We also exploit the redundancy in the WWW to predict and visualize outliers.

3.4.1 Views of Person Information

In autovita, our goal is to provide a comprehensive 360 degree view for a person. The system classifies all information on the WWW for a person into various 'Views' of presentation. A 'view' for a person is any of the dimensions of interest in person related searches like - background information, events participated, organizational relations , colleagues, financial appearances on the WWW etc.

Like some existing search engines that cluster the documents, we did not consider a 'view' for a person to comprise of a cluster of all the search result documents related to the person. This is because, documents from the WWW individually do not act as complete dedicated resources for a person's information. The data in a document consists of a number of varied topics as well as independent events possibly related to different people, organizations etc. Given a person, events associated with him can be listed anywhere in a document in separate sentences. Hence, our characteriziation for the 'view summary' to consist of sentences instead of documents is justified.

In the current system we have used the following 5 views a. Background b. Events c. Colleagues d. Organizational relations e. Finances

Each view consists of a set of patterns that characterize the view. Many representations for patterns have been successfully used for information extraction (Muslea, 1999). In Autovita, a pattern primarily consists of Noun Phrases (NP), Verb Phrases(VP). The phrases can contain a named entity tag such as a 'Person' name or 'Organization' name etc. In Autovita, we use GATE to do the entity tagging. About 6 different tags are used in the process of annotation of the sentences with entities. A typical pattern is of the following representation

$$[NP_{name}][VP_{\{work,establish\}}][NP_{organization}]. \quad (3)$$

In this pattern $[NP_{name}]$ represents any Noun Phrase consisting of a name and $[VP_{\{work,establish\}}]$ represents any Verb phrase with possible head words of the phrase as work, establish. Our matching of words is guided by the similarity function described in Eq(2). We also support regular expressions in the patterns for representing repetitive occurences of the constituting entities. For e.g.

$$[NP_{name}][NP]^*[VP][NP_{organization}].$$

When constructing a 'view' for a person we first instantiate each of the patterns associated with the view and perform the matching. Instantiation of a typical pattern discussed in Eq(3) with "Raj Reddy" can be represented as shown in Eq(4).

$$[NP_{name:RajReddy}][VP_{\{work,establish\}}][NP_{Organization}] \quad (4)$$

The pattern sets are applied one at a time. All the patterns in a set are matched starting at the first word of the sentence. If more than one pattern matches, the one matching the longest segment is selected; if more than one pattern matches the longest segment, the first is taken. The process of matching is discussed in greater detail in the following sub-section.

3.4.2 Extracting a View Summary

Our intention is to provide the least possible redundant information in all the possible 'views' for a person. Given the view, some patterns in sentences such as combinations of query words, named entities and phrases, may contain more important and relevant information than single words. Our redundancy removal algorithm makes use of the same. The goal is for the user to quickly get useful information without going through a lot of redundant information, which is a tedious and time-consuming task.

Our approach for picking the most relevant and non-redundant sentences that constitute a 'View' is motivated by novelty detection algorithms used in summarization techniques like (Zhang et al., 2002). In our approach, we use Maximum marginal revelance (MMR) with an adapted similarity calculation metric to detect redundancy. We believe that patterns such as named entities, phrases and etc, contain more important and relevant information than single words given a users

request or information need. Since our data is annotated with this rich information, we now use it in creating a 'view'.

MMR was introduced by Carbonell and Goldstein (Carbonell and Goldstein, 1998), which was used for reducing redundancy while maintaining query relevance in document reranking and text summarization. MMR starts with the same initial sentences ranking used in other baselines and our approaches. In MMR, the first sentence is always novel and ranked top in novelty ranking. All other sentences are selected according their MMR scores. One sentence is selected and put into the ranking list of novelty sentences at a time. MMR scores are recalculated for all unselected sentences once a sentence is selected. The process stops until all sentences in the initial ranking list are selected. In Autovita, for creating a non-redundant 'View' for a person , we use MMR which is calculated by Eq. (5). The similarity between the patterns in the 'View' and sentence is calculated as shown in Eq (6) and the similarity between the any two sentences is calculated as shown in Eq (8)

$$MMR = argmax_{S_i \in R/N}[\lambda(Sim_1(S_i, Vk) - (1-\lambda)max_{S_j \in N} Sim_2(S_i, S_j))] \quad (5)$$

$$Sim_1(S_i, V_k) = \sum_{ne_i \in V_k} \sum_{ne_j \in S_i} g(ne_i, ne_j) \quad (6)$$

$$g(w_1, w_2) \begin{cases} 1 & \text{if } w_1 = w_2 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$Sim_2(S_i, S_j) = \sum_{w_i \in S_i} \sum_{w_j \in S_2} sim_{fp}(w_i, w_j) \quad (8)$$

where V_k is the view in consideration for the person, S_i , and S_j are the ith and jth sentences in the initial sentence ranking. Q represents the query, N is the set of sentences that have been currently selected by MMR and R/N is the set of sentences have not yet selected. Sim1 is the similarity metric between sentence and query used in sentence retrieval and Sim2 can be the same as Sim1 or a different similarity metric between sentences.

3.5 Outlier Visualization

Another interesting aspect of the web is its redundancy. Information in repositories like the web is often redundant, in the sense that can be found in different contexts and in different superfcial formats. The redundancy of information can be a weak proof of its validity (Dingli et al., 2003). Our belief is that we could use the repetitive occurences on the WWW as a feature to score a web document which can act as credibility of the document for visitors visting the pages related to some particular person. In this manner, outstanding outliers in information and their source documents can easily be identified. In Autovita, we define an outlier to be any such sentence or information that does not come from an authorized set of webpages or is not repeated on different websites. We use the knowledge of link structure to identify an authorized set of web sites for the particular user. An authorized set of websites for a person are the homepage and all links initiating from the homepage or all those pages which are present in the same domain as that of the person's homepage.

Identification of outliers helps the owner of the content or actual person in context by altering him and helping him identify those pieces of information which probably have a negative connotation to them or are miscommunicated over the web. The assumption is that such outliers are not repeated on the WWW and the information on the homepages of the person is valid.

The system first picks all the sentences in a particular 'view' created for a person and the corresponding named entites. Each sentence is compared to the other sentences to identify the matching set of named entities. Based on the similarity obtained we identify those sentences that have non-repetitive information and also originate from unauthorized set of webpages . System then highlights all such extracted sentences and also provides a link to the source document from which this information was extracted.

4 Discussion of Experiments and Results

An experiment was conducted to demonstrate the system and its usefulness. The experiment involved automatically constructing a dossier on a given subject from information appearing on web pages indexed by Google. Information related to the subject was compiled using our Autovita system. Human review was then conducted to assess the kinds of errors found.

The subject on which the experiment was conducted was Raj Reddy, a distinguished computer

scientist and a Turing Award recipient. Entering Raj Reddy into Google generated 372,000 hits. The first 1000 text pages were selected and the information surrounding the occurrence of his name was extracted and catalogued. Human review of the material was conducted. The same set of results were fed into the Autovita system. The system could identify all the webpages related to only 5 of the 7 namesakes occuring in the top 1000 pages for "Raj Reddy". A lot of significant facts for each of the views were identified and a summary was generated. (Sweeney, 2005) lists some of the issues like conflict-coexistence, false positives, closed world distortion and inflated corroboration that are often the problems in web appearances of persons on the Internet. The results discussed in this paper, with the same subject were identified by the Autovita system. Where as the results in (Sweeney, 2005) were obtained semi-automatically and mostly by human annotation of the pages, we could reproduce the results in a largely automated manner.

5 Conclusion and Future Work

In this paper, we have discussed our system Autovita which mines person information from the web to provide comprehensive 360 degree 'view' for person related searches. We used a real world search engine results for a person's name and performed simple text mining and information extraction techniques to build a comprehensive view of the person. We proposed a technique that uses the homepage of a person for disambiguating the webpages of a person from among an enormous list of results returned by a search engine for a person name. We also proposed and used an adapted similarity metric that makes use of named entity tags to effectively extract a set of non-redundant sentences related to a person that constitute a 'view' for the person. Finally we used the frequency of web appearances along with the web page link structure information to predict outliers in information extracted.

Our system performs well and extracts information when the content of the webpage could be parsed to extract sentences. However, most web pages have a lot of information defined in the structure of the web page wrapped in HTML or other markup language tags and are not present as regular english sentences. In future, we would like to improve our information extraction algorithms for such web pages by experimenting with approaches like induction of wrappers (Kushmerick et al., 1997) and conditional random fields (McCallum, 2003). We would also like to experiment approaches for extraction of facts from unannotated data similar to the one proposed in (Brin, 1998). Integration of facts from multiple resources needs to be improved with the support of linguistics and inferecing mechanisms. In the current system, we have extensively used similarity metrics in our algorithms. We would like to explore more similarity algorithms and evaluate to pick a better one using tools like SimMetric library [4].

References

V.O. Mittal A.L. Berger. 2000. Ocelot: A system for summarizing web pages. In *Proc. of 23rd Annual Conf. on Research and Development in Information Retrieval (ACM SIGIR)*, pages 144–151.

Ron Bekkerman and Andrew McCallum. 2005. Disambiguating web appearances of people in a social network. In *Proceedings of the WWW 2005*.

Sergey Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of International Workshop on The World Wide Web and Databases at 6th International Conference on Extending Database Technology, EDBT'98*, pages 172–183.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Alexiei Dingli, Fabio Ciravegna, David Guthrie, and Yorick Wilks. 2003. Mining web sites using adaptive information extraction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: Is more always better? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*.

[4] http://www.dcs.shef.ac.uk/ sam/stringmetrics.html

Line Eikvil. 1999. Information extraction from world wide web - a survey. Technical Report 945, Norweigan Computing Center.

O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2004. Web-scale information extraction in knowitall. In *Proceedings of the WWW 2004*.

Alexiei Dingli Fabio Ciravegna, Sam Chapman and Yorick Wilks. 2004. Learning to harvest information for the semantic web. In *Proceedings of the 1st European Semantic Web Symposium*.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *SCIE*, pages 10–27.

N. Kushmerick, D. Weld, and R. Doorenbos. 1997. Wrapper induction for information extraction. In *Proceedings of IJCAI-97*.

I. Mani and M.T. Maybury (editors). 1999. *Advances in Automatic Text Summarization*. MIT Press.

A. McCallum. 2003. Efficiently inducing features or conditional random fields. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*.

Tom Mitchell. 2001. Extracting targeted data from the web. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining San Francisco, California*.

I Muslea. 1999. Extraction patterns for information extraction tasks: A survey. In *Proceedings of AAAI Workshop on Machine Learning for Information Extraction*.

Dragomir Radev, Tim Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Elliott Drabek, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. Mead - a platform for multidocument multilingual text summarization. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

L. Sweeney. 2005. Privacy-enhanced linking. *ACM SIGKDD Explorations*, 7(2).

Y. Zhang, J. Callan, and T. Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proc. ACM SIGIR 2002*, pages 81–88.