University of Waterloo at TREC 2011: A Social Networking Approach to the Legal Learning Track

Robert Warren
David R. Cheriton School of Computer Science, University of Waterloo

Abstract

This paper reports on the University of Waterloo experience with the Legal Learning track where three different methods were used to approach the retrieval task. Two are based on previously used methods and the last is a novel method based on modifying the responsiveness probability using social network analysis.

1 Introduction

The goal of the experiments this year was the exploration of whether social network analysis could be applied to the problem of e-discovery and legal document retrieval. We also opted for a fully automatic approach in that only responsiveness judgments from the topic authorities were used in the learning component of the system.

2 Run Descriptions

The Waterloo team submitted three runs to the legal learning track: a BM25 driven run [4], a machine learning run and a BM25 driven run modified according to a social network analysis probability function. In all three cases, the initial 'startup' run used was a BM25 retrieval run without pseudo-relevance feedback. The Wumpus Search Engine[1] which was developed by Stefan Buettcher while at the University of Waterloo was used to perform the experiments.

2.1 BM25 Relevance Feedback run (BAS)

In this run, the documents judged as responsive by the topic authorities were fed to the information retrieval engine using relevance feedback in the manner of Bodo [1]. At each iteration of the learning cycle, the top-100 most ranked un-accessed documents were selected to be send to the Topic Authority for evaluation.

The BM25 seed queries for each topic were:

401	enrononline online service
402	derivatives financial instruments
403	emissions spills pollution noise animal habitat environment

Table 1: BM25 query terms used as a 'startup run'.

2.2 Logistic Regression run (LRN)

A second reference run was created based on the same machine learning mechanisms used to recommend documents for manual assessment in the TREC-2010 UW submission [5]. Instead of using accessors, the results of the BM25 'startup' run and the answers of the topic authority are used to train a logistic regression classifier which then assigns probabilities to all documents within the collection. Because of the need for both responsive and non-responsive examples for the classifier, no initial run is presented for the logistic regression run.

2.3 Social Network Analysis run (SNA)

Our choice of a probability function for the SNA-based retrieval run was based on two considerations: 1) the function needed to be intuitive and model a specific social construct and 2) focus on information unlikely to be considered by the BM25 function.

We choose to model the probability that the sender of the communication is sending a responsive document. This prior probability is based on the number of responsive documents already sent by the person against the total number of their communications.

[1] http://www.wumpus-search.org/

$$P_{\text{doc}} = P_{\text{bm25}} * \text{AVG}(\forall P_{\text{doc}}(\text{sender})) \qquad (1)$$

This is used as a multiplier against the BM25 probability score as in Equation (1)[2]. As such, the addition of the term does not raise the probability of any document but instead lowers the probability of documents which have no evidence of responsiveness. We review this method in depth in Section 4.

3 Performance evaluation

For the preliminary evaluation, we provide comparative statistics on the hypothetical F1 scores for topics 403 and 401, the results of topic 402 being consistent across all three methods. A difficulty is that we used the basic run results to suggest documents to the topic authority which may have limited the learning aspect of our social networking method and hence its preliminary evaluation.

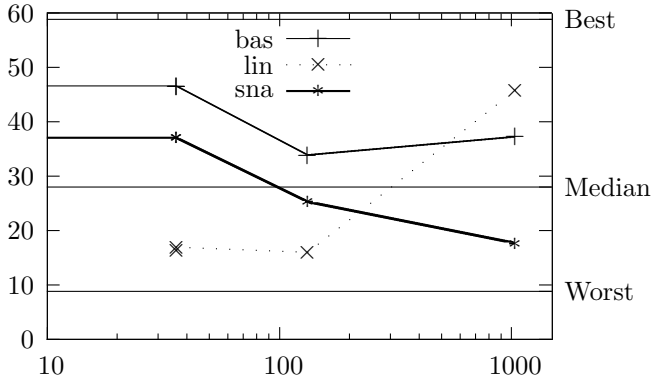

Figure 1: Incremental performance results on all runs on topic 401 against the number of responsive documents obtained from topic authority. The last data point consists of the mop-up run.

The logistic classifier generally did poorly for small amounts of training documents but its performance in mop-up runs were comparable to all other methods. Both the basic and sna methods were able to effectively function with few positive responses from the topic authority given their reliance on an initial start-up query.

Two interesting cases are plotted in Figures 1 and 2. In the first case, the social networking probability function diverged from the basic run in almost

[2]The equation as typeset in the original notebook paper was incorrect.

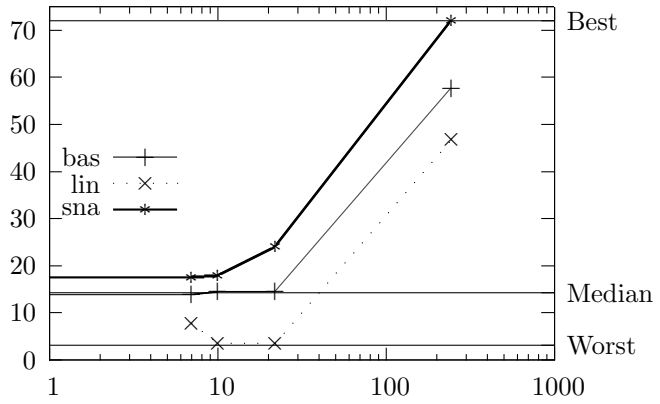

Figure 2: Incremental performance results on all runs on topic 403 against the number of responsive documents obtained from topic authority. The last data point consists of the mop-up run.

a mirror-like fashion against the median line. Given the large number of senders and of responsive documents, the social networking probability function aggressively selected from a new set of un-accessed documents. Our preliminary review of these documents suggests that these are indeed responsive documents and further evaluation of these methods is needed.

In the second case of Figure 2, the social network probability function was able to outperform all other teams for this specific topic. Analysis of the topic spurned several hypothesis which isolated topic 403 from the others: its topic being a non-core concern of Enron, the relative number of Enron employees versus non-employees and the specialization of the employees concerned with the topic.

As a topic dedicated to the environment, topic 403 is not a core concern for a company focused on energy markets such as Enron. It was hypothesized that this would have focused the discussion to a small group of people in Enron, facilitating an SNA approach. However, a second round of tests on the TREC 2010 topics did not support this hypothesis.

Similarly, it could be that the topic would be only of interest to a subset of Enron employees or that a small world effect would be occurring. Analysis of the senders, receivers and network revealed no significant difference from other topics, discounting this analysis.

The last hypothesis, that the topic is primarily an internal matter for Enron, proved to be true. All other topics had a large number of documents sent or received by people outside of Enron which skewed the normalization since not all of their communica-

tions were known. Topic 403 is a topic primarily discussed between Enron employees, permitting a probability normalization superior to that of all other topics. This serves to explain the ability of the sna driven method to outperform all other methods in this particular instance.

4 Discussion

The objective of the research was the application of a social networking approach to the discovery process. Popularized by the well know experiment by Milgram[6], some of the methods have been previously been applied to analyzing the Enron data, notably by Diesner [2] and McCallum [3]. We counted 255,964 individuals that sent or received documents within the dataset, which is close to the totals reported by other researchers accounting for variations on the parsing of email headers.

While the e-discovery field likes to refer to documents as belonging to custodians, we revert to the notion of sender and receiver of the document within the email system. This also makes the documents easier to model within the contexts of social networks.

Much of the work done in the area has been on the basis of graph theory, anecdotal observation, scenario analysis and of comparative metrics between datasets. For example, Diesner [2] reported that the Enron social network was more centralized and connected during the crisis rather than after.

A problem is that the pre-processing of the information can be tedious, influences the end results and the interpretation of the results is not always something that can be mechanized for use within a retrieval algorithm.

Figure 3 represents one of the basic elements of social network analysis in our context: How many communications is a person receiving 'in-degree' and how many communications is a person sending 'out-degree'.

These two simple concepts are key to our analysis of the dataset in that they are extremely simple, but can model multiple social constructs of interest. Take the example of a telephone network: a phone number with a high in-degree would be termed a 'pizza number' in that everyone calls a pizza delivery store but its popularity indicates that the communication has little actual value.

Similarly in the case of a company-wide email, the large audience (high out-degree), the wide dissemination of the information also indicates that the value of the information is low.

Given our requirements for a probability function

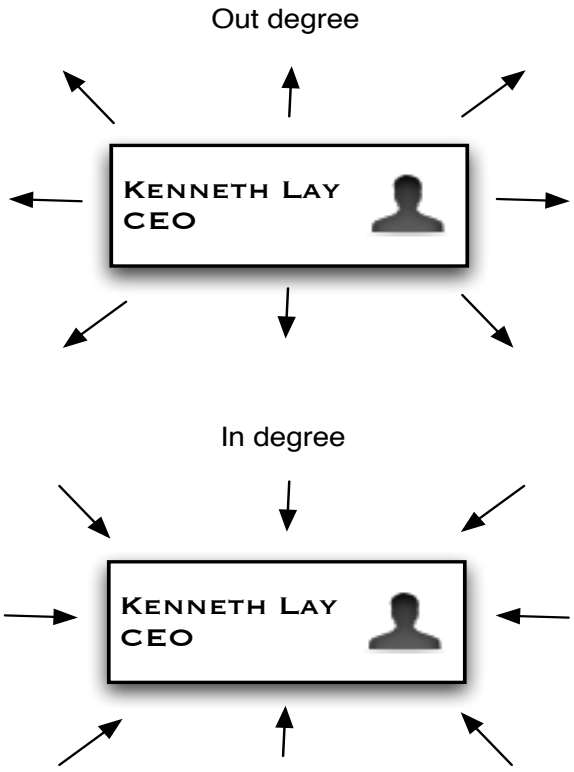

Figure 3: The distribution of a communication and the

that derives some information from the social network, we chose the form listed in Equation (1) which is based on the out-degree of a sender's communication. In effect, the probability function models the following theory: "a person whose work role involves communicating responsive documents will tend to communicate more responsive documents".

A criticism of this methodology is that it is not true Social Network Analysis: the social and audience context is not taken into account by the new probability term. This is true, but it is balanced by the fact that these network effects are implicitly taken into account by the BM25 function due to the duplication of messages and any other additions would results in breaking our second design objective.

What is clear is that the SNA probability function tends to select a different subset of the documents that have not been classified yet. In hindsight, we believe that the method would have done better had it been allowed to submit its own documents to the topic authority. Limited initial review of the documents selected by the function suggested that a high number of these were actually responsive but unclassified as of yet.

5 Conclusion and future work

In this instance of the TREC Legal Learning Track the University of Waterloo submitted three different automatic runs to the competition. We did find that well the three methods performed generally well, some where able to train using less training data than others.

A consideration is the erratic behavior of the SNA driven approach to retrieval. While the method does perform well in some circumstances it remains susceptible to noise within the dataset. The hypothesis at this point is that time is another dimension that the SNA formula must take into account.

In future work, the SNA scoring formula will depreciate the weight of relationships which are not maintained by new communications. This novel attempt at pruning the background noise should present a better quality of working set with which to score documents.

References

[1] Bodo Billerbeck. *Efficient Query Expansion*. PhD thesis, RMIT University, 2005.

[2] Jana Diesner and Kathleen M Carley. Exploration of communication networks from the enron email corpus. In *Proceedings of the Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining*, pages 3–14, Newport Beach, CA, April 2005.

[3] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email. In *Proceedings of the Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining*, pages 33–42, Newport Beach, CA, April 2005.

[4] S. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *NIST Special Publication 500-226:Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, 1992.

[5] Mark D. Smucker, Charles L. A. Clarke, Gordon V. Cormack, and Olga Vechtomova. University of waterloo at trec 2010: Legal interactive. In *Online Proceedings of TREC 2010*, number SP 500-294, Gaithersburg, MD, USA, 2010. National Institute of Standards and Technology, National Institute of Standards and Technology.

[6] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.