Expert Search using Internal Corporate Blogs

Pranam Kolari[†] pranam@yahoo-inc.com
Tim Finin[‡] finin@umbc.edu
Kelly Lyons[◇] kelly.lyons@utoronto.ca
Yelena Yesha[‡] yeyesha@umbc.edu

[†]Yahoo! Search Sciences, Santa Clara, CA 95054, USA
[‡]Department of CSEE, University of Maryland, Baltimore County, Baltimore, MD 21250, USA
[◇]Department of Information Studies, University of Toronto, Toronto, ON M5S 1C5, Canada

ABSTRACT

Weblogs, or blogs enable a new form of communication on the Internet. In this paper, we discuss blogs within a large corporation, and show their potential as a source of evidence to the expert search task. We describe characteristics of such blogs along multiple dimensions, and identify their utility to sub-problems within expert search. We finally discuss the use of blogs when combined with additional sources of information available within corporations.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications; H.4.1 [**Information Systems Applications**]: Office Automation

General Terms

Enterprise Blogs, Expert Search

Keywords

weblogs, corporate, enterprise, expertise, blogs

1. INTRODUCTION

Many of the challenges offered by information retrieval continue to fascinate researchers. One such challenge is that of identifying and ranking the creators of information, the problem of "expert search". The immediate importance of the problem to work-force efficiency, has clearly driven focused efforts within an organizational scope [7, 8, 14, 2, 13].

Given a topic of interest, the problem consists of three interrelated sub-tasks: (i) finding relevant and authoritative sources of information, (ii) identifying and associating individuals with this information (now evidence), and (iii) combining multiple such evidence to rank individuals (now experts). Any solution could leverage diverse information sources (e.g. documents, e-mails, wikis, and distribution-lists) hosted within an organization. Though subtasks (i) and (iii) are less coupled to the nature of content, (ii) is highly tied to it. For instance, in e-mail, the association problem takes a binary form, though less direct in other sources of evidence. This simplicity, and the potential to higher precision (clear association) and recall (organizational reach), motivated early research to explore e-mail as an important source of evidence [6].

However, a dependence on e-mail has limitations, the most important being that of privacy. Recognizing this inherent limitation, combined with the generalization (to the Web at large) enabled by the use of diverse information sources, the TREC (Text REtrieval Conference[1]) expert search tasks now promote and encourage the use of pubic facing organizational content [7, 1]. This typically includes distribution-lists and many other sources of publicly available content. With this as the background, it is evident that the research community continues to explore multiple diverse sources of evidence. Each such source supplements the other, with the eventual aim of increasing overall recall (and precision) in organizational expert search.

In this paper, we present one such additional source, internal organizational (corporate or enterprise) blogs. These encompass all non-public blogs hosted within the organization on their intranets. Employees use such blogs during the course of their daily responsibilities, to share information, voice opinions, protect ownership to ideas, and to initiate discussions on issues of general interest across the organization.

Our analysis is based on internal blogs (between November 2003 and August 2006) within IBM, a global technology corporation with over 300000 employees. Blogs are published using an extended version of Roller[2], an Apache powered open source platform. Each blog is owned by an employee, or a group of employees, with a total of around 23500 blogs. These blogs host 48500 posts with a similar number of comments. Posts carry with them a timestamp, author and tags that associate content to a folksonomy[3] of topics as perceived by the author. In addition, for this study, for every employee owning a blog, information on their geographical location, and to their position and chain in the corporate hierarchy is also available.

In complementing existing sources of expert evidence, blogs provide additional benefits: (i) unlike e-mail, available for expert search from the privacy perspective, (ii) unlike other sources, providing explicit author association, timestamp and metadata, in addition to (iii) hosting topically coherent snippets of information with implicit community vote through comments. An early evidence of reduced privacy concerns is evident from its availability for researchers like us, who are external to the organization. Looking forward, we believe that content within blogs has potential similar to e-mail, and can be viewed as a social bottom-up solution to separating out shareable content from non-sharable content within an organization. We revisit our earlier work [12] on the properties of internal corporate blogs to emphasize a few of these characteristics.

The rest of this paper is organized as follows. We first show that internal blogs provide a rich source of information by discussing their growth and content properties. We next detail network prop-

Future Challenges in Expertise Retrieval
SIGIR 2008 workshop, July 24, Singapore

[1]http://trec.nist.gov/
[2]http://rollerweblogger.org
[3]http://en.wikipedia.org/wiki/Folksonomy

erties and their implications. We finally discuss how certain unique characteristics of this content source could potentially enable new approaches to finding and ranking experts within an organization.

2. GROWTH AND CONTENT CHARACTERISTICS

Though less than 10% of the work-force engaged in blogs at the time of this study, their current growth suggests great long term potential. We also discuss the nature of this content, to validate how it could serve as evidence for expertise validation.

2.1 Growth of Users

At the time when it was actively tracked, the external blogosphere doubled every six months [15]. Internal blogs double at a little less than a year. Figure 1 shows the number of blogs and posts on a cumulative scale. The divergence between blogs and posts shows an interesting trend on how the blogging community is better engaging new adopters, and encouraging them to post content, hence retaining them.

To better understand how the creation of new blogs and posts trend over time we also plot the number of blogs created per month in figure 1. Two distinct spikes characterize this growth. The first, early in January 2004 was around the time when internal blogs were initiated within the organization. However, the second sharp rise around April or May 2005 was critical to the growth of blogs for two significant reasons, (i) the period following this is characterized by a dramatic increase in blog posts, and (ii) number of new blogs created every month has doubled from 500 to 1000 from before to after, suggesting that adoption was catalyzed. It turns out that at this time the organization officially embraced blogging as a communication medium and formally specified its policy and guidelines for both internal and external blogs. Evidently, having formal policies and a top-down guidance embracing blogs is key to the adoption of blogs by employees.

Driven by these organizational policy changes and high retention rate [12], we believe that the adoption is headed for continued growth, more so as the Facebook and Myspace generation enters the corporate world. Though we do not claim that blogs will support expert search task by itself, we believe its size will be significant enough to be a very important source of evidence.

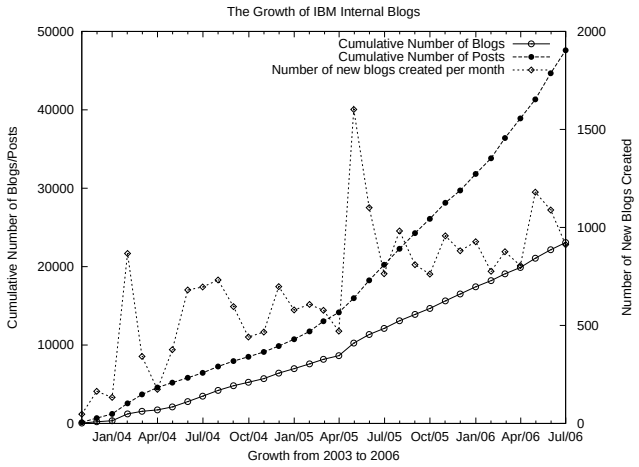

Figure 1: Growth of blogs and hosted posts has been phenomenal, with the number of blogs doubling every 10 months.

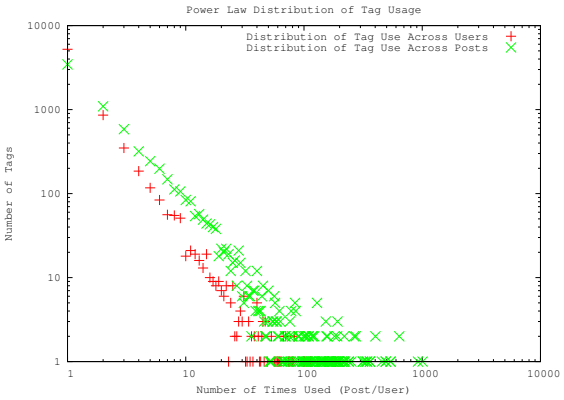

Figure 2: The distribution of tags based on their occurrence across all posts, and by the number of authors using them.

2.2 Discussed Themes

We next identify themes commonly discussed within internal blogs. We use the log-likelihood approach to compare language (word usage) distributions. Informally, this measure provides a profile of content genres. We compare a random sample of content in internal blogs with that of external blogs. We first list the terms representative of internal blogs:

> *IBM, Java, code, software, team, Microsoft, Sametime, Lotus, Dogear, innovation, client, wiki, collaboration, management, social, services, customer, support, products, Websphere*

Topics of organizational nature including products, competitors and work-enviroment related issues are widely discussed in internal blogs. In contrast terms representative of external blogs are shown below:

> *journal, she, her, me, him, love, girl, lol, god, im, mom, school, shit, night, gonna, friend, tonight, eat, cry, guy, sick, happy*

Clearly, external blogs feature day-to-day activities while internal blogs focus on themes important to an organization. Many of these themes are topics typical of organizational expert search, suggesting that content in blogs can be a useful source of evidence.

2.3 Use of Tags

Tagging is fast becoming a common way of associating keywords (tags) to organize content. The set of all tags within a specific system or application defines a folksonomy i.e. a vocabulary of terms. We analyze to see how tags, and the concept of folksonomy is adopted by blog authors. Close to 80% of all posts are tagged, with an average of over two tags per post. However, a discussion of folksonomies is incomplete without understanding its quality [11] i.e. a folksonomy is of very little use if every user uses a distinct vocabulary of terms.

We study two attributes that have a potential bearing on quality, (i) the distribution of tags across all posts, and (ii) the distribution of tags across users making these posts. A tag provides better value to a folksonomy when used many times, and by multiple users. Figure 2 shows the distribution of tags across all posts on a logarithmic scale. The usage follows a power-law distribution indicating that a small number of tags are used with a high frequency, and a large number of them are rarely used. Similarly, the second plot in figure

2 represents the distribution of tags across users. More authors using the same tags could potentially reflect well on quality.

Since tags are less susceptible to spam in a controlled enterprise environment, the general agreement on a subset of tags suggests many common and important themes are discussed across internal blogs, again providing additional content for expert search. Though arguable that tags might add little additional information gleanable from posts [4], they can still be used as a summarization of topicality, a key attribute for expertise evidence.

2.4 Links from posts

Using posts from 2 months, we analyze how many posts feature out-links (hyperlinks), both internal and external to the organization. 60% of all posts feature out-links of one form or the other. Out of these posts, close to 70% had links to the domain of the enterprise, 50% to other domains and 22% to other internal blogs. Clearly this data point further emphasizes that *employees largely blog about themes of interest to the organization they work for*. The use of blogs, as a complementary data source, can provide useful information on authoritativeness of other sources of evidence i.e. documents discussing topics of interest to the organization that are not necessarily blogs. These characteristics could be useful in evaluating the value of new expertise evidence.

3. NETWORK CHARACTERISTICS

We next move to the study of network properties of internal blogs. Many such properties have been found useful in expert search. To materialize a social network, we generate a directed graph $G(V, E)$, where V is a non-empty finite set of vertices or nodes, and E is a finite set of edges between them. Every user u, independent of whether she owns a single blog or multiple blogs, represents a vertex in G. A directed edge e from node u to node v exists in G, if user u has commented on, or linked to, a blog post made by user v. Each such edge represents an *interaction*. We call such a graph, a *blog interaction graph*, since it reflects interactions across users through blogs. G represents a social network across all users.

We pre-process G to eliminate self-loops, to collapse multiple edges between nodes into a single edge, and to prune disconnected nodes. After pre-processing, the graph consists of 4500 nodes with 17500 edges. In the rest of this section we discuss some of the structural properties of this network, and its implications to expert search. Our analysis makes use of the JUNG[4] toolkit.

3.1 Degree Distribution

The degree distribution of a network is significant in understanding the dynamics of a network and its resilience to the deletion of nodes [3]. For every node u in G, the in-degree d_{in} and the out-degree d_{out} is computed as the number of incoming and outgoing edges respectively. The in-degree $P(d_{in})$, and out-degree distributions $P(d_{out})$ are then plotted on a log-log scale, and the power-law exponents γ_{in} and γ_{out} computed using a line fit.

The in-degree and out-degree distribution of G follows a power-law as shown in figure 3, with $\gamma_{in} = -1.6$ and $\gamma_{out} = -1.9$. This is a little lesser than their values found on the Web ($\gamma_{out} = -2.67, \gamma_{in} = -2.1$) [5], but comparable to e-mail networks ($\gamma_{out} = -2.03, \gamma_{in} = -1.49$) [9]. In the context of expert search, this scale-free property of the network shows the *resilience of the community to user attrition*. It also shows how the network of users is amenable to finding experts.

[4]http://jung.sourceforge.net/

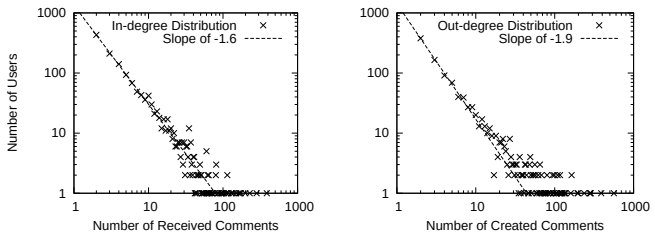

Figure 3: The in-degree of the network follows a power-law with slope -1.6 i.e. a few users generate most of the conversation. The out-degree of the network similarly follows a power-law with slope -1.9 with a few users contributing to many conversations.

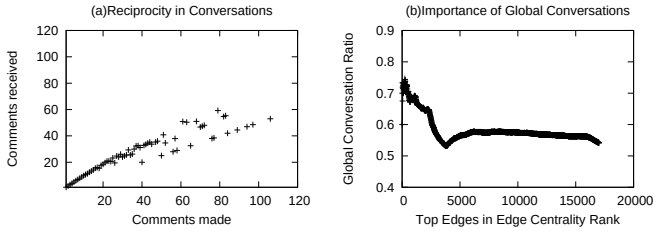

Figure 4: (a) A high correlation between in-degree and out-degree shows the reciprocal nature of blog comments. (b) A high number of cross-geography comments among high ranked central edges shows that blogs could help surface cross-geography experts.

3.2 Degree Correlation

Another interesting property of any communication medium is degree correlation. In blogs, it measures the reciprocal nature of comments i.e. *Do users who receive a number of comments, make a similar number of comments?* We plot the average out-degree of all nodes with the same in-degree. Results are shown in figure 4(a). The correlation holds for smaller degrees, but diverges randomly at higher values, possibly due to insufficient data points at such values. In general, active users in the community in addition to hosting comments on their own blog, also contribute to comments on other blogs. This has an interesting implication. It suggests that many of the popular authors (if experts) are also keen in engaging with other users within the organization. These are the experts within the organizations who could be more receptive to queries from other individuals.

3.3 Edge Betweenness Centrality

Betweenness centrality [10] measures the significance of nodes and edges as it relates to their centrality in information flow through the network. It hence forms an important measure for identifying effective word of mouth channels within a community. Many of the central nodes are key connectors within the organizations. To identify if edges that reflect interactions across geographies are central to the network, we rank edges based on their centrality, computed by finding the number of times a specific edge features in a shortest-path between every pair of nodes.

Using a ranked list of such central edges, we plot the distribution of edges that cross geographical boundaries (countries). As seen in figure 4 (b), the high ratio of such cross-geography edges

among the top ranks show the value of global interactions. Such edges form significant bridges to information dissemination across a global organization. Unlike other sources, blogs are known to be more effective in surfacing experts from such interactions.

Readers interested in many other related network properties, including those of graph ranking, are referred to our earlier work [12] on corporate blogs.

4. DISCUSSION

Motivated by many of these properties, we developed an expert search prototype for use within the organization. The application used a simple approach to topical expert search, (i) posts that serve as evidence on a topic were identified through matching tags (ii) such posts were associated to their unique authors, and (iii) these authors were ranked for expertise by comparing the aggregate number of comments to all their posts (a simplistic voting model). The tool was exposed as a tag cloud, with topical expert search limited to tags in the folksonomy. The developed application was submitted to IBM's internal hack day (which required the application be developed within a day), and was voted among top five entries by IBM employees. It was also showcased through other initiatives within IBM. Many of these resulted in feedback that can provide interesting cues to expert search, moving further. We discuss one such direction that involved tuning the comment based voting model.

In a conversation (all comments around a blog post), the relative position of employees part of the interaction, as measured through the corporate hierarchy, can be useful to understand the reach and spread of posts, and in-turn topical expertise. To support this, the employee hierarchy is modeled as a rooted named unordered tree, T. The root of the tree is the head of the organization. Each employee-manager relation is represented using a parent-child relation making managers internal nodes in the tree, and all non-managerial employees leaves.

We briefly introduce some basic tree properties. A node is an ancestor of another node u, if it exists in the shortest path from u to the root node. The height of a node u in T, denoted as $h(u, T)$ is the distance between the node u to the root of the tree, with the height of root node being zero. The Lowest Common Ancestor (LCA) of any two nodes u and v in a tree is the lowest node in T that has both u and v as descendents. We define a sub-tree $T^{u,v}_{LCA}$, as a tree rooted at the LCA of u and v and featuring only nodes and edges that are in the path from u and v to the LCA. $E(T)$ is the set of all edges in the tree T.

As opposed to only using the number of comments, the concept of *spread*, defined as the number of edges in the union of all comments around a blog post could be useful. Spread is defined as:

$$S_p(u, V) = \frac{|\bigcup_{v \in V} E(T^{u,v}_{LCA})|}{|V|}$$

Noticeably, the distribution of normalized spread across all blog posts peaks at around four [12], suggesting that conversations are high across users working in close hierarchical proximity, and less exclusive among peers, and between employees and their managers. Overall, we believe this property of conversations could signify an interesting attribute of blogs. If e-mail conversations are evidence of expertise from a 'peer' perspective, and generic documents (or mailing-lists) are evidence from a 'global' organization perspective, blogs could potentially be evidence from the 'department' as a whole.

Though the above hypothesis demands further analysis, it does point to an interesting new direction to quantify the utility of blogs. More generally, it suggests reviewing existing sources of evidence within expert search, and evaluating and accommodating new sources. While we begin to extend this work to the more generic expert search, we encourage researchers to continue exploring blogs as a useful source of evidence.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions made by Stephen Perelgut and Jennifer Hawkins at IBM Toronto Software Labs, and IBM for supporting this research.

6. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2006. ACM.

[3] A.-L. Barabasi. Emergence of Scaling in Random Networks. *Handbook of Graphs and Networks*, pages 69–84, 2004.

[4] B. Berendt and C. Hanser. Tags are not Metadata, but Just More Content - to Some People. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[5] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[6] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531, New York, NY, USA, 2003. ACM.

[7] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC-2005*, Gaithersburg, Maryland USA, November 2005.

[8] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@NOPTIC Expert: Searching for experts not just for documents. In *Poster Proceedings of AusWeb(2001)*, 2001.

[9] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66, 2002.

[10] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[11] M. Guy and E. Tonkin. Folksonomie, tidying up tags? *D-Lib Magazine*, 62(1), 2006.

[12] P. Kolari, T. Finin, Y. Yesha, Y. Yesha, K. Lyons, S. Perelgut, and J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[13] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *30th European Conference on Information Retrieval (ECIR 2008)*, 2008.

[14] D. Mattox, M. T. Maybury, and D. Morey. Enterprise expert and knowledge discovery. In *Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction*, pages 303–307, Mahwah, NJ, USA, 1999. Lawrence Erlbaum Associates, Inc.

[15] D. Sifry. State of the blogosphere, october, 2006.