NetPosse: A Tool for Connecting Users in Virtual Communities

Faisal Anwar
Edlab at Teachers College, Columbia University
New York, NY 10027 (USA)

Hui Soo Chae
Edlab at Teachers College, Columbia University
New York, NY 10027 (USA)

Gary Natriello
Edlab at Teachers College, Columbia University
New York, NY 10027 (USA)

ABSTRACT

We discuss the design of Netposse, a tool that matches people in online communities based on their intellectual and professional interests. We frame the motivation of this tool around current research on how learners can leverage social communities for maximum benefit. Given this literature, we present a design for Netposse that mines data from existing web assets and matches people according to their areas of expertise. In addition to matching individuals, such a system serves the purpose of query answering as well: it allows users to search and identify others in the community whose background qualifies them to provide advice on a user's topic of interest.

1. INTRODUCTION AND BACKGROUND

NetPosse is an intelligent online tool that enables users of online communities to efficiently query one another for information and support. NetPosse seeks to do for person to person communication what traditional search engines have done to connect individuals to web content. As online systems mature, virtual communities such as those on the Internet are not just hosting documents and media, but individuals as well. Consequently, there are significant opportunities to create tools that connect people to one another as they spend more of their lives online.

The most direct pattern of use for the NetPosse framework is as a tool that connects novices and experts within a community of practice [6]. In this case, a novice is anyone searching online for help and information on a particular topic. While online media and documents may be useful in some contexts, the novice may find that talking to an expert in his community is the most effective way to master the topic at hand. In such a situation, NetPosse's role is to connect the novice to experts within his community who would be of greatest help. NetPosse's scope also extends beyond connecting novices to experts in a community. It can be an equally effective tool for connecting two experts who may want to discuss a difficult problem together as colleagues.

In addition to facilitating learning interactions among individuals with different degrees of expertise, NetPosse has the potential to lower the transaction costs and thereby improve learning outcomes among individuals who are only weakly connected. Exposure to newer ideas comes from interaction with those with whom we are weakly tied, because such individuals travel in different social circles [7] and thus have access to information and resources that we do not [4]. Yet, those to whom we are only weakly connected are less motivated to share this information as compared to strongly tied individuals whom we trust and work with closely [5]. By facilitating learning interactions among those who are weakly connected NetPosse can activate a more complete learning matrix.

2. NETPOSSE'S DOMAIN OF USE

In the current context, we limit the discussion of NetPosse to a specific domain of use: connecting students and faculty at Teachers College (TC), Columbia University.

The diagram below outlines the basic virtual structure of the TC community. Teachers College has several different web assets and several different populations (e.g., students, faculty and alumni) that it serves. In terms of web assets, there are likely several different sites such as the main college website, the alumni website, and a site for the libraries. There may also be overlap between information sources on the wider Internet and within TC. For example, there are Facebook communities that exist specifically for TC students.

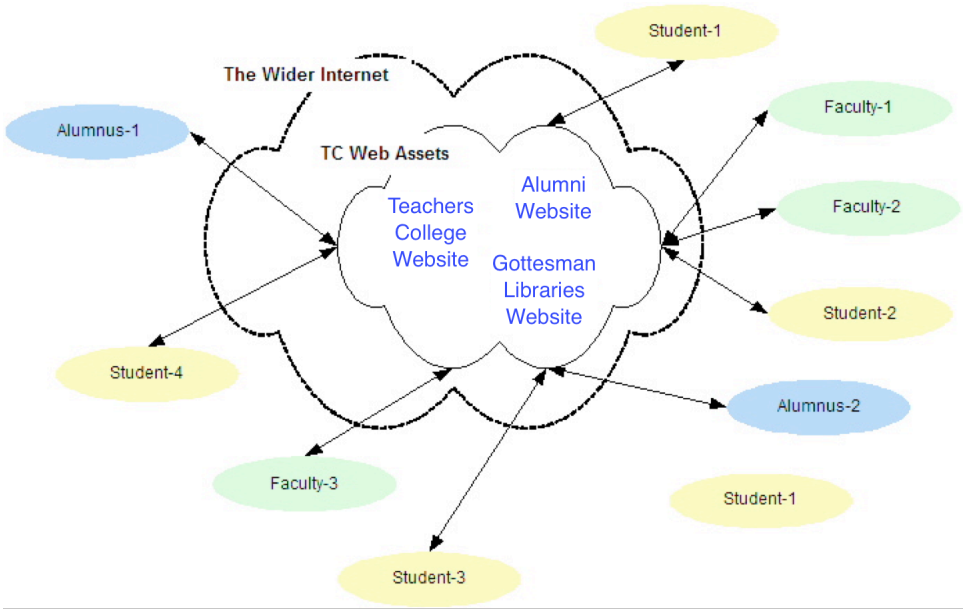

Figure 1: Outline of Teachers College's web topology.

Given this community structure, let's suppose that a particular student, Jane, is doing some background research for her doctoral thesis. To find out more about her topic, she can pursue some typical research strategies:

- She can access information physically or virtually through TC's library system.
- Her advisor can probably point her to some promising avenues for investigation. The advisor may even be able to suggest specific people that she should talk to for expert advice on her question.
- She can rely on personal contacts which may include students, faculty members and alumni to get more information on her topic.

The goal of NetPosse is to complement and enhance Jane's research experience by identifying experts in the TC community with the content knowledge and willingness to help Jane with her query. In Jane's situation, Netposse will use an intelligent matching algorithm to quickly and conveniently identify who might benefit Jane the most in her research. This strategy builds upon similar approaches for people matching, such as the data-mining technique employed in discussion forums by Dringus and Ellis [3]. Moreover, NetPosse is intended to support queries of multiple levels of complexity – from simple questions about TC's administrative policies to deeper questions related to academic research.

3. TECHNICAL OUTLINE OF THE NETPOSSE FRAMEWORK

There are a range of issues NetPosse technology must address if it is to be a compelling and relevant tool for community members. Figure 2 below outlines the key modules that will work together to create the final NetPosse user experience.

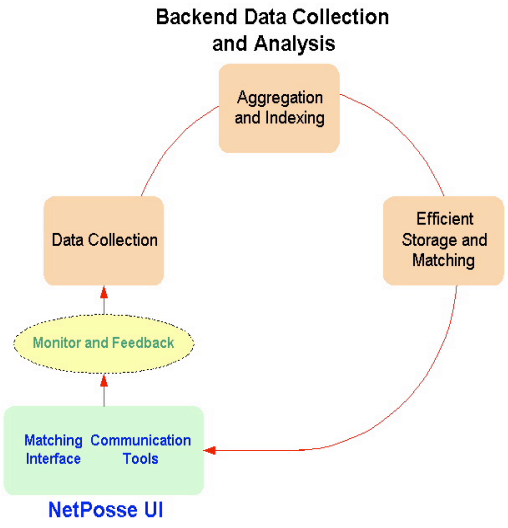

Figure 2: NetPosse Workflow

As the diagram illustrates, NetPosse requires a significant investment in backend technologies that support the final user experience of finding an expert on a given topic. Data collection, aggregation and indexing are needed to create a knowledge base. Using this data, an efficient matching algorithm can perform the task of answering queries posed by novices by identifying helpful experts available in the community. There are several key questions that must be addressed by any technical implementation of the data collection, aggregation and indexing stages:

- Which sources of data are used to inform the matching algorithm and how will these sources be accessed for crawling and indexing?
- How will community members provide access to biographical data that will be used to construct the knowledge base?
- While accessing different sources of data, how are issues of privacy and security handled? How do community members know about and control the information that is used and/or disseminated by NetPosse? Alternatively, can NetPosse connect groups of people to other groups where the individual members maintain a degree of anonymity unless they volunteer to share their identity?
- When data from multiple sources is aggregated for a particular community member, is there any reasonable strategy for prioritizing the data according to quality and relevance? That is, are all sources equally important in identifying a person's area of expertise, or should some types of sources be given higher priority?

Once data has been collected and stored, it will be used by a matching algorithm to efficiently identify community members who can best assist a novice user with his query. In addition to relevance, algorithm performance is an important consideration since NetPosse will compete against traditional content-based search engines that have quick response times. Finally, NetPosse's matching process must ensure that users are not so overburdened by queries from novices that they disassociate themselves from the community.

The key to avoiding over burdening people would seem to lie in developing precise searches that limit exposure to questions squarely within a target person's domain of expertise. Even so, individuals in certain areas may still find the number of queries unmanageable, so there may need to be some provision for distributing popular queries to others with similar expertise, to authored documents, or to specially constructed faq files. Finally, future iterations of the system may give experts some control over the flow of queries that they receive both so that they might be more motivated to participate and so that we might monitor their use of controls and gain a greater understanding of preferences.

The NetPosse user interface consists of two key elements: a matching tool and a communication mechanism. The task of the former is to present a simple interface through which users submit queries and view community experts who may be of help to them. The communication piece of NetPosse handles how users get in touch with one another. To keep the interface as familiar as possible, we intend to provide directory entries as the output of a NetPosse search. In this way, the NetPosse user experience is very similar to that of a search engine: a user inputs a query of interest and the system outputs a list of choices. In NetPosse's case, the choices are individuals (and how to get in touch with them) as opposed to website names and hyperlinks. However, it is important to note that we are designing NetPosse so that user communication can be extended to other digital communication paradigms such as email, Facebook, and Twitter. The focus of NetPosse is to match people within a community – once such matches have been made, the communication paradigm should be extensible to many different contexts.

Finally, we also intend to implement a feedback loop for our system once a basic prototype has been constructed. The purpose of the feedback loop is to monitor how users employ NetPosse and to adjust the data collection and matching processes accordingly. We expect that NetPosse will eventually become an adaptive matching framework that presents users with experts on a given query based on insight from previous NetPosse sessions.

4. DESIGN OF THE NETPOSSE SEARCH FACILITY

We now present the design and rationale of the core search modules of NetPosse. These are represented by the 'Data Collection', 'Aggregation and Indexing', and 'Efficient Storage and Matching' boxes in figure 2. The basic problem these modules attempt to solve is to take an input query (in the form of search terms) and provide a list of users who have knowledge or expertise about that query.

Data collection

The most important task in the search process is identifying the data sources that will be used to build knowledge about a user's community. NetPosse will function within the context of an existing user community, and the data sources will come from the online assets of that community. In the case of Teachers College, for example, we can identify several data sources:

- The PocketKnowledge archiving system [9].
- Online course management systems such as Live Syllabus [1].
- Public websites and information pages.

- Published work in journals, conference proceedings, and other accessible publication venues.

Data from each source will be crawled by NetPosse on a regular basis, akin to how websites are crawled by search engines. In this case, the data collection process will center on individuals as opposed to websites. Figure 3 outlines the basic process that we are implementing.

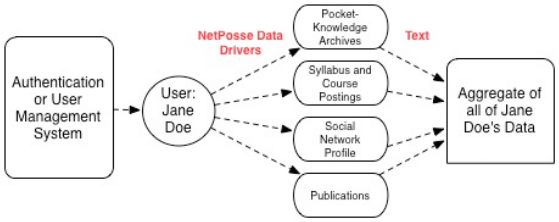

Figure 3: Data collection strategy.

The critical difference between the traditional search paradigms is that we are now indexing around individuals rather than websites. This indexing occurs within the context of some online community, so that instead of crawling a wholly public space of web pages, we must provide a scheme for crawling available information for one person at a time.

We will use a bridge service around a community's user directory or authentication system to identify all users of interest. For each user, NetPosse will then crawl online resources to build a body of text produced or actively consumed by the user. NetPosse will need to access systems that go beyond standard web pages. To support this, we will implement individual NetPosse data drivers for each data source. These data drivers will be customized to each possible data source and will output user-specific data in a standard format that is then aggregated within NetPosse. The final output of the data collection module will be a text-based archive for each user gleaned from all available data sources in his community. After some further processing, this data will serve as the knowledge base for fulfilling queries for experts on different topics.

Aggregation and indexing

The next task at hand is to take our aggregated user archives and build an index or model that will be applied to queries. The basic problem is to take a body of text representing a user's knowledge and create a model that can be easily searched and matched to queries. We take an approach that has some important distinctions from standard indexing strategies.

Our model is based on multinomial text classification systems similar to those outlined by McCallum and Nigam [8]. In particular, each person represents a different class and the multinomial model for that person is constructed using the text that is collected earlier from data sources. Assuming we have N words in our multinomial vocabulary, our index then is a column vector of length N representing the frequency counts of each word in the vocabulary. We will use a condensed vocabulary in which the words included are those that have the greatest information value, as recommended by McCallum and Nigam.

Matching strategy

Once we have generated a multinomial model for each community member, we can then use these models to answer queries. Each query will be submitted as a set of words that a user wants information on. Instead of matching to documents, NetPosse will now match to people, identifying those individuals who match best to the query. We can use the multinomial counts of each word to help us score different people as matches. For each person in the community, we add up the total multinomial counts for each word in the user query. These total counts then represent the overall score for each community member for the query provided. Then, NetPosse outputs the users according to who received the highest scores among all users in the system.

5. CHALLENGES AND FURTHER AREAS OF RESEARCH

Presently, we are actively developing the NetPosse system and have identified several areas for further research. First, we would like to iteratively improve our search algorithm. We have based the initial design on a standard multinomial indexing scheme. Such a design allows us to fulfill short search queries while giving us the flexibility to match on larger bodies of text at a later time. However, we would like to test how effective this scheme is in a real user community and to improve the design further. One possible refinement would be to add a weighting scheme related to the recency of text preparation so that queries to experts are directed to their more recent work. Such an approach might help manage the volume of queries received by any one expert and provide a greater motivation for experts to respond based on the assumption that they would be most interested in questions on current interests.

Secondly, we acknowledge that there are some challenges in providing the right incentives for community members to share information. Many community members may avoid NetPosse participation because of privacy issues. To allay these concerns, we are initially limiting the data collection process to publicly available information. However, we believe that personal information such as courses taken or personal profiles on social networking sites will be important in creating accurate models of people. We are therefore designing a strategy where users can opt to provide this information anonymously in a way where it is only used to improve the search, but not available in any way for other users to view.

6. CONCLUSION

This proposal lays the groundwork for a new person-to-person matching technology called NetPosse. The goal of NetPosse is to help people identify and virtually communicate with others who are most able to help with a particular query. Fundamentally, NetPosse works by building a relevant knowledge base about the people in a community and then employing a matching algorithm that will match queries to the data in the knowledge base. NetPosse works analogously to document search engines, except that it organizes crawling, indexing and matching tasks around people rather than websites. This creates new requirements for the architecture and implementation of the system. In particular, we have outlined a person-centric data aggregation and indexing scheme that will be used to build knowledge to answer user queries.

7. REFERENCES

[1] Chae, H. S., Mitchinson, A., Dai, H., Sathe, P., Anwar, F., & Yuan, T. (2009). Live Syllabus: Building an Intelligent Networked Course Syllabus Tool. In *TCETC Conference Proceedings*. New York, NY.

[2] Cocciolo, A., Chae, H. & Natriello, G. (2007). Using Social Network Analysis to Highlight an Emerging Online Community of Practice. Paper presented at the Conference on Computed Supported Collaborative Learning. New Brunswick, NJ, July.

[3] Dringus, L. P. and T. Ellis (2005). "Using data mining as a strategy for assessing asynchronous discussion forums." *Computer & Education Journal, 45*, 141-160.

[4] Granovetter, M.S. (1973) The strength of weak ties. *American Journal of Sociology, 78*, 1360-80.

[5] Haythornthwaite, C. (2002). Building Social Networks Via Computer Networks: Creating and Sustaining Distributed Learning Communities. In *Building Virtual Communities: Learning and Change in Cyberspace*, edited by K.A. Renninger and W. Shumar, pp. 159-90. Cambridge, UK: Cambridge University Press.

[6] Lave, J. and E. Wenger (1991). *Situation learning: Legitimate Peripheral Participation*. Cambridge, UK, Cambridge University Press.

[7] Kadushin, C. (1966). The friends and supporters of psychotherapy: On social circles in urban life. *American Sociological Review, 31*, 786-802

[8] McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*. Penn State University.

[9] Mentor, M., Strome, E., Asunka, S., Agnitti, G., & Natriello, G. (2008). Early returns on an institutional repository: an exploration of the validity and functionality of Pocketknowledge. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital LIbraries* (p. 459). Pittsburgh, PA: ACM/IEEE-CS. Retrieved May 19, 2009, from http://portal.acm.org/citation.cfm?id=1378889.1379004.