CS 294 Jonathan Huang
Human Centered Computing Nov. 30, 1999
Final Report

A Study on the Effective Usage of Social and Knowledge Structures in a Note-based Document Sharing System

Introduction

The purpose of this project is to investigate the potentiality of applying principles of centrality and prestige as seen from the context of social and knowledge structures to a set of collective information. In this particular project, this collective information is a set of more than 5000 notes taken by individuals in digital format that resides in a central repository as part of the NotePals system. The idea behind this system is to allow its users to record meeting records in a digital format on devices such as a Personal Digital Assistant or a Crosspad, and later upload these documents to a central repository that is viewable online via an web-based application. The inspiration behind this system is that during a meeting and other collaborative settings there is usually a lot of contextual information that is shared by participants of the discussion. Since its participants are constantly transforming this information, it would be inadequate for one individual to act as a squire to try to capture all the information that is available. Instead, the approach is to collect the notes taken by each of the participants and later combine these notes into a coherent format. The system currently allows individuals to search and access these shared documents by specifying parameters such as a document's organization hierarchy within the system, its author, its type, the date it was taken, and/or any keyword(s) it may contain. Although this method of searching through this knowledge base may seem valid from a database point of view, it does not fully exploit the implicit social structures that may be embedded between these different documents. This project investigates how users of this system can

exploit the ideas of centrality and prestige to increase the likeliness of finding the desired document(s) and also finding related documents that may be of additional use to them.

Centrality

There are many interpretations of centrality in the context of shared documents in a note system. At the simplest level, the *actor (as discussed in Wasserman & Faust's Social Network Analysis)* is the individual who has taken the most notes given a particular filtering criteria. For example, an individual with 20 notes on a particular subject as opposed to another who has only 5 notes on the same subject is the more central actor since more documents are attached to him as the base. For this interpretation of centrality the statistics relating documents to authors can be easily calculated and stored for later reference. To take this process one step further, this information can be processed while the documents are being uploaded and added to a lookup table for future reference.

Another way to determine centrality is to cluster the access pattern of notes by query and reduce it down to relationships between individuals and documents. For example, given a query the system can track the notes the user visits and record the access statistics for later lookup. Presumably for notes that are accessed most often one would assume that there is an implicit relationship between the search criteria and the viewed notes. However, this approach is not as systematic and sound as the previous method in that it relies a lot on inferring the user's intent in looking at those notes, and it can be easily skewed by a wondering individual who is just browsing the notes in the system. Nonetheless, it would be helpful to provide users with statistics on the notes that are accessed most often with a particular criterion as an additional method for filtering through notes.

Prestige

In the context of note-taking prestige is an indication of an individual or group's authority on the content of one or more notes. These individuals do not necessary have to be the author of these notes, but could also appear as references within the note documents. This is analogous to the way authors are cited within conference papers and scientific journals, where authors that are cited most often are usually those that are distinguished in a particular area. For the note system to automatically pick up references to individuals, it can rely on handwriting recognition to build a list of commonly referenced names and organizations. The system would then use these authority rankings to order the notes it returns by prestige, so that the user can choose to look at the notes from a more "distinguished" source first. This system of organizing notes by prestige can also help users find related notes. This is under the assumption that once the user has found an author who has a high ranking of prestige in a particular area, other notes written by or related to that author may also be of interest to the user. Using these embedded reference users can browse notes by following references rather than perform a rudimentary search using metadata information.

On a looser level, prestige can also be determined by finding the relationship between the quantity of notes and their authors. Although this approach is valid from the point of view of centrality, this inference is much more precarious for determining prestige based on notes. Whereas a large number of notes is usually a good indication of author centrality to a particular area, it may or may not be an indication that an author is also an authority on the subject. Furthermore, this approach assumes that respected authors also uses this system often and has many notes related to his field of study. Nevertheless, the system could display this information

in conjunction with the ranked authors to give the users a better idea of which authors are active note takers in their respective areas and which are well known (cited) in the collection of notes.

Collaborative filtering is yet another method for determining author prestige within this context. For this approach users would rank the usefulness of the notes they've visited in terms of content, and the system can then keep track of these rankings to create an index of authors and subjects. This approach cannot be automated by the system because it relies on the interpretation of the notes in order to determine its quality. However, by including the user in the feedback loop the system can easily retrieve this ranked information and provide it to the user as a relation between authors and their authorities on particular subject(s).

Another way to determine prestige within this context is to assign status to the individual users of the system, provided that they provide some background information about themselves. However, there are many levels of categorization and ranking possible for individuals both within and between groups, and sometimes it is difficult to distinguish between who has more authority on a particular subject for individuals belongs to different groups. For individuals within a particular group, say individuals related to Computer Science ranked by the number of years of education they have, it is most likely that notes taken by a CS Professor will carry more weight than an undergraduate CS student. For the between-group case where two individuals are from different group and the subject matter is related that of one group, one can assume that the individual from that related group should be able to provide better notes than the one from the other group. However, when individuals from different groups are compared to a third (neutral) subject, this ranking of prestige or expertise is not as obvious. An example might be a Computer Science versus an Architecture student taking notes for a biology class. In this scenario it is difficult to distinguish who is more closely related to the subject at hand, but a possible solution

may be to rank the majors with respect to closeness to each other (e.g. Computer Science is more closely related to Math than Biology). Therefore, by employing a hierarchy of related fields and years of education it is possible to determine prestige of authors given the notes taken by a particular individual.

The above cases seem pretty straightforward, even for the case where two individuals from different groups are compared to a neutral subject area. The complexity, however, comes from the fact that individuals often belong to multiple groups of practice, and the problem of determining authority based subject closeness now becomes unclear. One approach might be to assign values to the different closeness rating between subjects and sum these values for different individuals, such that the one with the bigger sum in the more prestigious of the two. However, this approach seems impractical considering an individual who is in CS with closeness rating five compared to an individual belonging to six different groups with each of closeness rating one will seem more prestigious in this method. A possible heuristic for determining prestige in this situation is as follows: if any of the individuals have a direct tie to the subject in question then the one with the more related subject is the more prestigious. However, if individuals need to go through more than one path to get to the subject in question, then the individual with the shortest path but largest closeness factor will be deemed the more prestigious of the two.

Proposed application

I am proposing to add centrality and prestige as additional filter options when looking for a particular set of notes. These two criteria can be used to order notes in a way such that a particular group of notes or one of more individuals becomes the focus of the search and thus relevant information surfaces. This can easily be done by having these two options as radio

buttons along with an additional “neutral” button as the third option. There should also be a streamlined process such that for a given query a user can use the information retrieved by ordering by centrality or prestige to deduce the author that is of interest and from him find related documents. Furthermore, the backend will need to be modified to record access statistics as well as record additional user information such as the fields they belong to and their relative experience in that field(s).

Survey

For the survey we are interested in finding out: 1) whether users understand the model of ordering by centrality and prestige; 2) why do they share notes between groups; and 3) what kind of additional organization structures they might want to apply to these documents. For this project I’ve conducted a survey study of 15 individuals made up of students from UC Berkeley, and we’ve found that notes on average do exhibit a high degree of centrality and prestige. For centrality, notes are clustered heavily around areas such as lectures, seminars, projects, and web-related documents. As for prestige, it is not surprising to find professors as the dominant source of information, followed by teaching assistants, members of a research community, and finally friends and classmates. Furthermore, the participants have ranked authority as the most effective method of searching for notes, followed collaborative filtering, and clustering based on access. This is not surprising since college student usually regard professors are their primary source of information. If professors were not directly available, then they would refer to their peers for sources of information. Access patterns of notes rely more heavily on personal preferences, which could be an indication why the participants of the survey did not regard it as an effective method of retrieving information.

Conclusion

Social and knowledge structures are important in information retrieval because as humans the documented knowledge that we create usually cannot exist without an implicit social context. Current search-related engines do not fully exploit this relationship in their criteria when letting users find documents of a particular nature. This paper has provided an analysis of the different ways that centrality and prestige can be incorporated into a search-based collaborative knowledge base and has proposed methods for doing so. The surveys indicate that users like the idea of ranking and ordering documents according to authority and centrality and would prefer using such a system over standard search methods using metadata. Furthermore, they find the ability to find related notes via authors of authority to be especially valuable in the newly proposed system. Therefore, it is clear that it is beneficial to bring the social context into this note-taking system to provide better ways for individuals to find the desired notes.