CS497 Information Retrieval
Project Proposal

A Personal Academic Event Manager (PAEM)
– Information Extraction Tool for Semi-structured Texts

March 11, 2003
Liping Chen(lchen2@uiuc.edu)
Yan Shi ( yanshi@uiuc.edu)

1 Introduction
We propose building a software tool called Personal Academic Event Manager (PAEM) as an implementation of information extraction (IE) technique in the domain of seminar announcements and call for papers. We regard documents from these domains as semi-structured data since documents of the same domain share common information structures and they are not entirely free texts. The same technique, however, could be easily applied to other domains such as organizational activities (meetings, appointments). The emphasis, though, is on adapting the existing IE algorithm to our specific domain rather than on fancy GUI design or comprehensiveness of the tool's functionality.

2 Our tool and potential users
As an active participant in the realm of research, we receive many seminar announcements or call for papers everyday coming from various sources. We might get overwhelmed by the large amount of these data and as a result, miss important conference or seminar information even if we are interested. Thus there is a demand for a tool to automatically process the incoming data and provide managing and searching tools. In this project, we would like to build such a tool called Personal Academic Event Manager (PAEM) focusing on two domains: seminar announcements and call for papers.

3 Previous information extraction tools built
After doing a search on Google, we find some programs built for extracting information from html based website or off-line data. But none of these software tools are very influential. And for most, we don’t know what they are capable of doing or how good they can do what they claim to do. However, our motivation for working on our version of the tool is that we think if we concentrate on semi-structured text instead of any text source, which is

1
Figure 1: Three layers to build the tool

more of the case in the domain we have defined, we could achieve a better performance.

Some of the results found:
Sophia: http://www.celi.it/english/sophia.htm
IntelliSonar: http://www.quigoint.com/about.htm

4 How to build it?
We consider building such a tool consisting of three layers: the underlying IE layer, the functionality layer, the GUI layer which is shown in Fig 1.

In this project, those domains we choose to work on are seminar announcements, call for papers etc. With the help of the inherent semi-structure of documents in these domains, we are to build scenario templates. For instance, in seminar announcements, we are interested in such slots as: title of the talk, speaker, location, date and the time. Thus the main task of IE layer is to extract the relevant information and perform the slot filling. We are now surveying one possible way of IE using maximum entropy classifier. Once the
proper information has been extracted for each slot, we could start to build functionality layer on top of the extracted data. Simple functions such as searching and scheduling based on time could be performed. More functions could be added if time permits. The GUI layer is built hand in hand with the functionality layer. It should facilitate the use of those functions provided by the underlying layer.

5 What to deliver?

It really depends on how suitable the existing IE algorithms are. We can’t promise anything definite at this point in time. However, we plan to do the following: 1. If the existing IE algorithm gives rather satisfactory result, we would like to devote more time on the functionality layer to provide more comprehensive tools. 2. Otherwise, our main focus would be to explore various existing IE algorithms and try to derive our own.

6 Timeline

We plan to do this project in two phases: from now to the end of March, we are to explore the maximum entropy classifier and build the IE layer. In the April, our main task is to develop the application layer and GUI layer. Along the way, each phase should be properly documented so that we could have our report handy when we finish the implementation. As said before, the time proportion might be adjusted along the way depending on how well the IE layer could be implemented.

References

[1] Hai Leong Chieu & Hwee Tou Ng A maximum Entropy Approach to information Extraction from semi-structured and Free Text, 2002