

What follows is an email exchange I had with a C.S./GSLIS/GenomicBio professor that will be of interest to CAPT. He betrays some naive ivory tower perspective on how easy it should be to implement things, not understanding the constraints of commercial, closed source, consortially deployed production systems. Also, our recent prioritization exercise has relegated some of these things to the bottom of the pile. Nevertheless, there could be some good ideas that come from CAPT explorations with him. Particularly the next-generation OPAC investigations seem ripe with opportunity.

As you can see from my last message to him, I'd like to suggest this be a topic for the upcoming CAPT meeting, at least enough to gauge interest.

On Oct. 12, ChengXiang Zhai wrote:

Quote:

Subject: Collaboration on deploying new information management technologies

From: ChengXiang Zhai <czhai@cs.uiuc.edu>

To: libsys@uiuc.edu

Dear Jodi,

As I discussed with you over the phone, I'm very interested in exploring opportunities for deploying some advanced information management technologies that we are developing in my research group through our library system. Could you please help me get in touch with the relevant people in our library?

You can find more information about me from my homepage at <http://www-faculty.cs.uiuc.edu/~czhai>.

Thank you very much!

Best,
-Cheng

On Thu, 12 Oct 2006, John Weible wrote:

Quote:

Hello Cheng,

Jodi forwarded your message to me. Can you describe more fully what you're asking of the Library? From your brief message, it's not at all clear what sort of collaboration you have in mind. By knowing more what you envision, I can hopefully either indicate a method to proceed or get you in contact with the right people.

Thanks,
- John Weible

On Oct. 12, Cheng wrote:

Quote:

Dear John,

Thank you for your reply!

Please allow me to introduce myself first:

My name is ChengXiang Zhai. I'm an assistant professor of Computer Science with a joint appointment in the Graduate School of Library and Information Science and the Institute for Genomic Biology. My main research area is information retrieval, though I have also been working on bioinformatics. I joined UIUC in 2002.

My research group have been working on a number of topics in the general area of text information management, including information retrieval models, personalized search, visualization of search results, collaborative and content-based filtering, text summarization, and subtopic pattern discovery/analysis from text. We have developed software components in all these areas, and we are very interested in developing and deploying some application systems using these component technologies. Since almost all these techniques can be potentially used to improve/extend our current library information systems, I naturally thought of exploring such collaboration opportunities.

Here are some specific thoughts:

* The current catalog search system seems to be not very accurate. For example, if I type in "information retrieval" or "information retrieval algorithms", the top ranked results are generally not very good. It seems to indicate that the underlying retrieval model/algorithm doesn't put more weight on the title field. The "Start of title" option is more accurate, but this is only useful if a customer knows the title.

We have developed some new retrieval models that have been shown to outperform most existing retrieval models in our experiments. We currently implemented our models on top of some existing open-source retrieval toolkits (e.g., Lemur/Indri and Lucene), and we can also implement it in our own library system, or if it's appropriate, we could also adopt one of these open-source retrieval toolkits which also have some other functions that we developed.

* We have developed effective techniques to personalize search results. The general idea is to use more information about a user to improve the search results for the particular user. For example, the past queries from the user, the past book records viewed by the user, and the books checked out by the user can

all be very useful to help improve the ranking of the search results for the user.

We have already implemented these techniques in a client-side Internet Explorer plug-in that can personalize Web search for a user. Evaluation indicates that our method can outperform Google by up to 20% in retrieval precision. We can port such techniques to our library system to improve search results for all the users.

* We also have techniques for recommending books to a user. The recommendation can be based on either content-matching (i.e., matching the book description with a user's interest which we can infer from the user's search history) or collaborative filtering (i.e., exploiting other similar users' preferences like what Amazon did), or both. With such techniques, when we get new books, the system can automatically recommend them to the users who are potentially interested in them.

* We have tools for discovering subtopics from a collection of text and analyzing the temporal trends of the topics. With such techniques, we can mine the publications in multiple fields to discover emerging interesting interdisciplinary topics. Such techniques can also be used to automatically generate a subtopic-based summary of all the proceedings/journals in some area. For example, we can process the ACM digital library and automatically generate subtopics in the computer science field and show which topics are getting more attention recently.

I also have some other ideas that can potentially improve the utility of our library system (e.g., we could aim at making our library website a major UIUC information service center where people can also socialize).

Anyway, I hope I've given enough information about what I have in mind. But I also realize that these may still too abstract. So perhaps a more productive way to proceed is for me to have a meeting with you, in which I can explain these to you in more detail and perhaps also show you some demo. What do you think?

Sorry for sending such a long message and thank you for taking time to read it.

Best,
-Cheng

On Oct. 17, John wrote:

Quote:

Hi Cheng,
We have a "Content Access Policy and Technology" committee at the Library, which is charged with oversight of the content access systems we use. All of the

types of things you propose fall into that group's purview, and some activities are underway in those areas.

The CAPT committee meets this week, and I will share with them your message. That way we can identify which people should be involved with the initial discussion with you and get back to you regarding scheduling.

- John