

CS598CXZ

Homework 1

Hieu K. Le
netID: hieule2

Web Domain: *Entity-Based Search*

As far as we know, Keyword-Based Search method has major drawback in expressing desired results. The ultimate goal of Information Retrieval is to acquire relevant information from different sources and organized them as structured was possible.

For data sources as web pages, being able to extract relevant information and put them into relation data is a significant challenging task. However, if we could do that, there are many applications which are very useful but can not be applied due to difficulty of working with web data become possible.

One attempt to reach this goal is to retrieve as much as information of attributes of an entity as possible given some known information about the entity and a specific Entity Relationship Diagram. This work has been started with relational data and was called as Entity Retrieval. Applying and extending the technique which was used in relational data to web data are also non-trivial task.

Users: Could be anyone who concerns about information any particular entity in a specific domain. For example, in Computer Community, people often want to find information about a particular researcher. User could do searching by give the kind of entity and value of some attribute that they know about that entity. The search result will be the values of attributes of that entity as complete as possible.

Data: Indexed web pages.

Challenges: There are many different variants of values of attributes of an entity in the real world. For example, with a researcher, his name could be cited in many different ways. To be able distinguish different entity given variants of some attribute is a very difficult problem.

Email Domain: *Automatically Mail Organizing*

The number of email each person receives per day is continuous increasing. There is a need of automatically organizing email in a reasonable and tractable ways.

One approach could be trying to connect emails which are related in some ways such as: discuss about the same problem, mention about the same event.

Like in a newsgroup, when a user wants to post a message, he will decide whether or not to start a new thread or will follow an existing thread. By doing this reasonably, the messages in the newsgroup will be much more easy to keep track with and to examine.

If we could do that for emails come to a mail box, this will make users manage email much more efficiently.

Users: Who have to contact by email frequently.

Data: Existing emails and new arriving emails.

Challenges: How to know an email is relevant to others in a specific aspect is a non-trivial task. High accuracy and explainable mechanism are required.

Literature Domain: *Automatically Discover Cause-Effect relationship.*

In literatures, facts as cause-effect relationship are popular, especially in medical, law, and history literature. To be able to do that, a person need to read all the related documents, remember most of facts, and do a good reasoning. However, with a huge number of literatures in each field today, no one could be able to do that thoroughly. Most of attempts success with some forms of lucky which is reaching right documents at right time.

Making this task done automatically, much useful and maybe surprised knowledge will not be missed. And base on this, we could build some a new kind of expert system which works directly with knowledge in form of literatures.

Users: Researchers, lawyers, historians.

Data: Existing literatures, especially literatures of medical, law, history, and chemistry.

Challenges: Recognizing and connecting causes and effects together is extremely hard.