Collective Resolution of Identity in Email Archives

Tamer Elsayed
Dept. Computer Science, University of Maryland
October 29, 2007

1. Problem

Access to historically significant email collections poses challenges that arise less often in personal collections. Searchers of such archives may need help making sense of (1) the identities of individuals that participated in the discussions observed in the collection, and (2) the terms exchanged in these discussions that may be too specific to the community of the participants. In this project, we will focus on the problem of resolving identity in the context of large email collections, leaving the dual problem of content resolution as a clear extension to it.

2. Solution Overview

By resolving identity, we aim to determine the true referent of a given reference or mention in any of the emails in the collection. Currently, our mention resolution approach makes use of the contextual space surrounding the mention of interest. Different types of context (e.g., topical) together comprise the contextual space of the queried mention. In each type of context, evidence is combined and thus candidates for the queried mention are ranked. The approach is currently applied to individual mention queries and runs in a single-thread process per query.

In the context of Hadoop, we propose two levels of extension to the current implementation:

1. We can design a MapReduce model of the single-query resolution technique. On one hand, this new design and implementation should make the resolution process much more efficient. On the other hand, larger contextual space can now be explored and probably contribute to a more effective resolution.

2. We can design a "collective (or joint) resolution" algorithm that aims to resolve all mentions in the email collection simultaneously. Such joint solution has not been computationally attractive unless a multi-process paradigm such as Hadoop is available.

3. Resources

In our experiments, we will use the Enron email collection [2]. Up to our knowledge, it is the largest real email collection available for research purposes. It includes 517,431 messages in 150 top-level directories; each of which contains the retained emails of a former Enron employee. The current resolution algorithm is written in Java using Lucene [1] as a back-end for indexing emails. A GUI interface for searching and exploring the collection is already available.

4. The MapReduce Perspective

For resolving a single mention, a mapper is supposed get a single email in a specific context and provides credits for candidates based on the relevant evidence observed in that email. A reducer will then simply sum the credit for a single candidate. If we consider the evidence that we currently leverage in resolving one mention, which is observed in the context of other mentions, we can envision the setup of the problem as a graph in which nodes are the mentions and the directed links represent the relative dependence between mentions in specific type of context, as shown in Fig 1. Such setup is very similar to the problem of ranking Web pages, to which PageRank [3] is the famous solution. We know that PageRank algorithm is a perfect fit of the MapReduce framework and we expect a similar design for our solution once we map the problem into the graph structure.
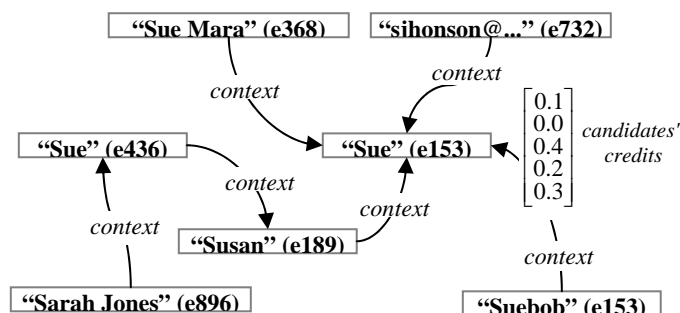

Fig 1. Modeling of collective resolution as a graph

5. Interesting Extensions

The dual problem of the identity resolution is what we call "content resolution". By resolving content, we seek to efficiently retrieve emails that are on-topic but may not match the query terms. For this problem, we propose a "document-expansion" approach that tries to reconstruct the context (e.g., by building large language models) and thus improve the perceived accuracy and utility of email retrieval task. MapReduce model can clearly be a useful framework here. We also hypothesize that there is a potential for an interplay of the solutions of the two main problems; resolution of identity and content. The joint resolution of all mentions will enrich the context of the mention, hence improving the content resolution.

6. REFERENCES

[1] http://lucene.apache.org.

[2] Bryan Klimt and Yiming Yang. Introducing the Enron corpus. In Conference on Email and Anti-Spam, Mountain view, CA, USA, July 30-31 2004.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Stanford technical report, 1998.