Sharing the Collective Intelligence between E-mail Applications

Boris Chidlovskii, Jean-Baptiste Faddoul
Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan, France
chidlovskii@xrce.xerox.com
jean-baptiste.faddoul@xrce.xerox.com

ABSTRACT
This position paper explores the idea of sharing the collective intelligence among different e-mail applications. The goal of such a sharing is to improve the performance of the e-mail services, like categorization, actor role and community detection on the large scale. We describe our approach to the cross-application categorization represented as the learning of multiple tasks with different category sets. Our multi-task learning approach is based on the ensemble principle and extends the conventional Adaboost technique from one- to multi-task setting.

1. INTRODUCTION
Advanced e-mail applications and services for users (actors) communicating by e-mails are expected to be non-invasive; moreover they will address domain modelling aspects, which are often specified in the form of user requirements or by a particular scenario. The most important parameters to take into account are the number of users and the specificity of tasks raised by the users. This latter may have different aspects. First, these tasks address the end user’s or administrator needs. Second, the variety and heterogeneity of activities represented in e-mails and how far the actors can go in sharing fragments of their e-mail space to accomplish the common tasks. First, it may depend on the mail-per-actor ratio; high values of this ratio point to cases when one person uses multiple e-mail addresses; low levels indicate multiple actors sharing the same address.

Below we consider three different scale cases, and discuss some e-mail-relevant services in the function of size of actors and their e-mail collection:

Individual mailboxes may contain thousands of mails and hundreds of addresses in their contact lists. Thus an individual user often expects a better organization of the contact lists, as well as such categorization of e-mails which would reflect various groups and activities (s)he participates in. Both standard and personalized services on grouping the contacts and categorizing the e-mails in users’ mailbox in the function of their main and secondary activities are expected to be an important step forward.

Small to middle size groups can be composed of two to one-two dozens co-workers in cooperative setting around one or few common activities. One example is the executive managers sharing personal Outlook mailbox with their assistants for preparing the agenda and tracking urgent actions. Another example is project-relevant or forum-based activities and actor communities, such as Xerox Codendi code sharing systems1 where groups of developers share their code and use e-mails for discussions, comments and meeting reports. These cases are grounded on the voluntary participation aimed at the common knowledge sharing. Management of the mail exchange within the group is built around the common project or few relevant topics and may include the project-relevant categories, mail deduplication and event tracking [6].

Large user-mail collections may contain thousands of users and billions of e-mails; dedicated decision-making tools become indispensable on the corporate and organization level. User (administrator) needs can vary according to the legislation in place. Apart from the standard detectors of the malicious email attacks in a large social context, these needs solicit novel mechanisms for tracking multiple actor communities, profiling multiple types of activities, topic and role discovery [5] in projects, events, personal communications, large scale broadcasting, etc.

Changing the scale may have an important impact on the goals and performance of the e-mail relevant services. Larger collections dispose much more data thus enabling the data mining techniques to run their exploration and discovery routines. On the other hand, advantages of smaller user-email pools are in a lower variety of activities; in most cases this makes the categorization problem easier.

1.1 E-mail categorization
Categories associated with e-mails reflect the variety of users’ communication activities, referring on the top level to categories of private communication (family, relatives and friends), personal (phone company, the bank, public services, etc.) or professional (co-workers and colleagues at work). Each of these top categories is often segmented in a number of more specific sub-categories.

Mail organization in users’ mailboxes is often sophisticated in terms of categories or topics, as far as an e-mail may belong to more than one category. Categorization of incoming and outgoing e-mails appears as an important service at both individual and large scale. Indeed (semi)-automatic categorization may facilitate the individual processing/archiving of e-mails or accelerate the domain expert search on the corporate scale. Multiple organizations

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
OrgCI’10 November 7, 2010, Florida, USA.
Copyright 2010 ACM X-XXXX-XX-X/XX/XX ...$10.00.

1http://www.codendi.com/
currently implement the e-mail categorization, by prompting the users to manually select the category of their e-mails from a limited list of pre-specified categories.

The manual categorization is expensive and the new trend in e-mail categorization systems consists in deploying the data mining and machine learning techniques. They count on advances in extracting specific features describing e-mails and learning models from available e-mails with categories. The methods commonly cope with two complementary sources of information, one describing the e-mail content and second describing the set of communicating actors, with the eventual goal of possibly combining the two sources in order to improve the global system performance.

While the e-mail content is recognized as the more informative source, identifying the specific social context of an actor in the network may play an important discriminative role for the categorization. The social network features are often capable to catch the position/role of e-mail actors (both sender and receivers) in the corporate network. It has been shown that certain properties of nodes in a communication graph can serve very well to automatically detect the social role of actors in the network [1].

1.2 Cross-application categorization

It became common that people participate in multiple social networks and communicate from multiple e-mail addresses. Sharing the collective knowledge between the applications represents an additional source for improving the categorization performance.

When sharing on the low scale, like between an individual’s home and office mailboxes, we face the case of complementarity where e-mails from different applications share some categories and may overlap in an important part, like in personal and private e-mails. The simple union of e-mail instances with categories often gives an immediate gain. The situation is different when sharing on the large scale, with thousands of different users and categories. If the large collections expose distinct category sets and hierarchy of roles played by corporate actors, we need to first mine both collections and establish points of similarity between them. For these challenges, we get help from the methods of transfer and multi-task learning.

2. MULTI-TASK LEARNING

The multi-task learning is a current trend in machine learning that attempts to go beyond the standard methodology of learning one task at a time. When a problem becomes too large, it is broken into small, reasonably independent sub-problems that are learned separately and then recombined in a proper way. The goal of multi-task and transfer is to improve the performance of related tasks by learning a model which is able to represent the common knowledge across tasks [2].

The current works on multi-task learning focus on neural networks, k-nearest neighbours, and support vector machines, where the common knowledge is explicitly expressed as a shared part of the tasks. Most of these techniques make some convenience assumptions, like one when the different tasks share the same topics/categories and examples. Moreover, they often make a hypothesis that tasks tend to behave similarly in the whole learning space.

The validation of the task relatedness represents a critical condition prior to any transfer between different applications and associated tasks. Unfortunately, dissimilar tasks might hurt the performance similarly to introducing noise in data and making global relatedness assumption turns to be too strong in real situations. Moreover this relatedness may show up different degrees or even different signs in different regions of the "learning space". It is therefore important that the multi-task learner determines the relatedness of tasks, learns its different degrees and accommodates the inductive bias accordingly.

2.1 Multi-task Adaboost

In order to achieve the mentioned goals, we develop an approach to the multi-task learning based on ensemble principle, in particular the boosting technique. Our multi-task learning approach, called MT-Adaboost [3], extends the conventional Adaboost algorithm [4] to the multi-task setting. It uses a new class of multi-task weak classifiers which are multi-task decision stumps (MT-Stump). An MT-stump is organized in multiple levels where each level is a decision stump for one task. Thus any MT-stump defines a partition of the learning space dependent on the “presence” of a given task in the given partition. Thus an ensemble of weighted multiple MT-stumps allow to learn the relatedness between tasks in different regions of the learning space. The main advantage of this approach is in relaxing the previous requirements on sharing labels/examples and global relatedness hypothesis.

The method has been first validated our approach on synthetic datasets. We are currently testing it on Enron and Tobacco datasets which are publicly available large scale e-mail collections with important textual and social aspects. The two collections have a comparable but different category sets. We probe different tasks from the two collections with MT-Adaboost and compare it to the state-of-the-art algorithms (SVM and Adaboost). The preliminary evaluation shows that MT-Adaboost can yield an important increase (2 to 8%) in the categorization accuracy.

The current multi-task prototype is being developed for various scenarios of e-mail mining in a collaborative environment, including the community detection and actor role discovery.

3. REFERENCES


