Collaborative OpenSocial Network Dataset based Email Ranking and Filtering

Ismael Rivera[1], Myriam Mencke[1], Juan Miguel Gomez[1], Giner Alor-Hernandez[2], Angel Garcia-Crespo[1]
[1]*Computer Science Department, Universidad Carlos III de Madrid*
[2]*Division of Research and Postgraduate Studies, Instituto Tecnologico de Orizaba, Mexico.*
[1]*{ismael.rivera, myriam.mencke, juanmiguel.gomez, angel.garcia}@uc3m.es*
[2]*gineralor@computacion.cs.cinvestav.mx*

Abstract

Social Networks have experienced a meteoric rise recently. They provide a number of functionalities such as network of friends or business contacts listings, content-sharing, profile surfing, discussion and messaging tools. Interoperability among Social Networks being a key challenge, the Google-powered OpenSocial alliance has partly solved it and unveiled a new breed of strategies to gather data from Social Network users. In this paper, we build on the OpenSocial functionality and combine it with filtering and ranking algorithm to enhance email management.

1. Introduction

Social Networks provide the means to explicitly create and manage connections based on information gathered and stored in user profiles. Social Networks (SN, in short) and Semantic Social Networks [18] have emerged as a second generation of the mailing lists, Usenet, bulletin boards online communities, providing a number of services such as network of friends or business contacts listings, content-sharing, profile surfing, discussion and messaging tools.

SN are also part of the recently created new breed of user generated content aware technologies which have been encompassed by the “Web 2.0” buzzword umbrella and have turned up to provide a huge amount of metadata and information about the user as a particular entity. Tags, picture sharing environments, social bookmarks, blogs and music preferences are just the top of the iceberg.

However, these applications are not addressing fundamental problems of information overload, such as email hoarding or lack of management, but contributing to increase the burden. On the other hand, efforts such as [16] and [17] are under way to examine email filtering and ranking based on social networks.

In addition, semantic technologies are evolving to a more mature state in which ontologies [1], its backbone technology, provide a formal representation of a domain. The shift enabled by the use of machine understandable ontologies can outperform the current endeavors that require finding data spread out across the Web or dynamically drawing inferences which are continually hampered by their reliance on ad-hoc data frameworks.

In this paper, we present a Google-powered OpenSocial based strategy to filter and rank email using SN user information. The contributions of the paper are twofold: firstly, we describe the OpenSocial Network Dataset (OSND), a lightweight ontology to garner SN user data, similar to [12]. Secondly, we couple the OSND with a number of information retrieval, filtering and ranking algorithms to enhance mail management. Ontology-guided Input tool, which provides query refinement and multifaceted browsing. We evaluate our proof-of-concept implementation.

The remainder of the paper is organized as follows. In section 2, we show our study case. In section 3, we discuss the OpenSocial Network Dataset as a basis for collaborative filtering [2]. In section 4, we describe our filtering and ranking algorithms to harness email overloads. In section 5, we show our future implementation and the first preliminary results of our choices for the prototype. Section 6 spans over and bind together a number of related works. Finally, section 7 concludes the paper and outlines our future work.

2. Scenario

The Internet provides several powerful tools and ways to work; email is one of them. Juan uses email for work, however, he also receives more than 250 emails a day and many are junk email messages from people he does not know. It would be extremely useful for him to be able to prioritize the emails from his professional contacts and most important friends, thus using his time more efficiently.

978-0-7695-3105-2/08 $25.00 © 2008 IEEE
DOI 10.1109/ICONS.2008.34

In addition, as more people use email, marketers are increasingly using email messages to pitch their products and services. Some consumers find unsolicited commercial email, also known as “spam”, annoying and time consuming; others have lost money by subscribing to fake offers that arrived in their email inbox [3].

But, how can we determine if an email message is from one of Juan’s friends or not? The answer is social networks.

Juan uses several of these networks. He is registered at Hi5 and Bebo where he has many friends and at LinkedIn, where he makes professional contacts. Hence, Juan will consider it more important to read the emails from Javier, a friend of his who is registered on Hi5, from Ismael, another executive from a company who is one of his contacts on LinkedIn and Myriam, a girl who he has met on Bebo.

3. OpenSocial Network Dataset

OpenSocial is an application programming interface to build social applications across the Web, in other words, a common set of APIs for social applications across multiple websites. With standard JavaScript and HTML, developers can create applications that access a social network's friends and update feeds [4].

OpenSocial is currently being developed by Google in conjunction with members of the web community. The ultimate goal is for any social website to be able to implement the APIs and host 3rd party social applications. There are many websites implementing OpenSocial, including Engage.com, Friendster, hi5, Hyves, imeem, LinkedIn, MySpace, Bebo, Ning, Oracle, orkut, Plaxo, Salesforce.com, Six Apart, Tianji, Viadeo, and XING [4].

OpenSocial is not a social network itself; rather it is a set of three common APIs that allow developers to access the following core functions and information on social networks:

- People and Friends data API: allows client applications to view and update People Profiles and Friend relationships using AtomPub GData APIs with a Google data schema. These applications can request a list of a user's Friends and query the content in an existing Profile.
- Activities data API: allows client applications to view and publish "actions" in the OpenSocial platform using AtomPub GData APIs with a Google data schema. This API allows the creation of new entries, editing or deletion of existing entries, and the capability to view lists of entries.
- Persistence data API: allows client applications to view and update key/value content using AtomPub GData APIs with a Google data schema. Applications can edit or delete content for an existing application, user, or gadget instance, and query the content in an existing feed.

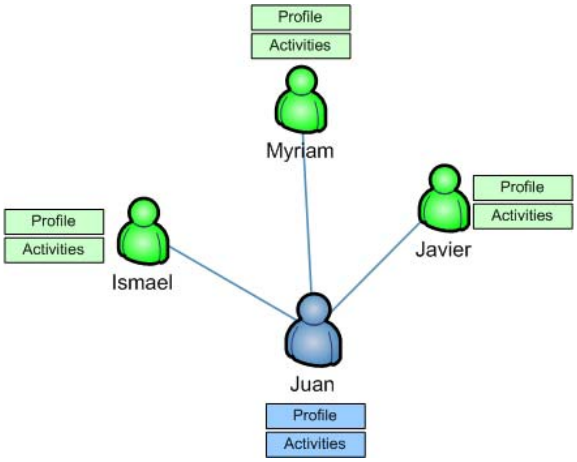


Fig. 1. Social Network of Juan

OSND is a lightweight ontology used for email ranking and filtering. It is constructed using the information from a set of social networks, getting a structured version of users’ profiles, getting a list of a user’s friends for each user and following their friend connections in order to get detailed profiles. We can obtain which people are friends of a user and how important or close they are.

Furthermore, the ontology uses the information about users’ actions, such as indicating when a user uploads a video file or a photo to a site, etc.

However, another fundamental feature is the possibility of tagging the content in all these applications. Tags are freely chosen keywords describing a particular resource. They offer a simple way of retrieving content (e.g. retrieval of my interesting communities in LinkedIn with the tag Semantics). These tag sets and their assignments to objects are envisaged as subjective conceptualizations, being potentially aggregated to a flat bottom-up categorization or folksonomy. In [5], Folksonomies have been claimed to be an interesting emergent attempt for information retrieval but serve different purposes to ontologies, the latter are attempts to more carefully define parts of the data world and to allow mappings and interactions between data held in different formats. Hence, ontologies are defined through a careful, explicit process that attempts to remove ambiguity, whereas the definition of a tag is a loose and implicit process where ambiguity might well remain. Finally, the inferential process applied to ontologies is logic based and uses operations such as join. The inferential process used on tags is statistical in nature and employs techniques such as clustering.

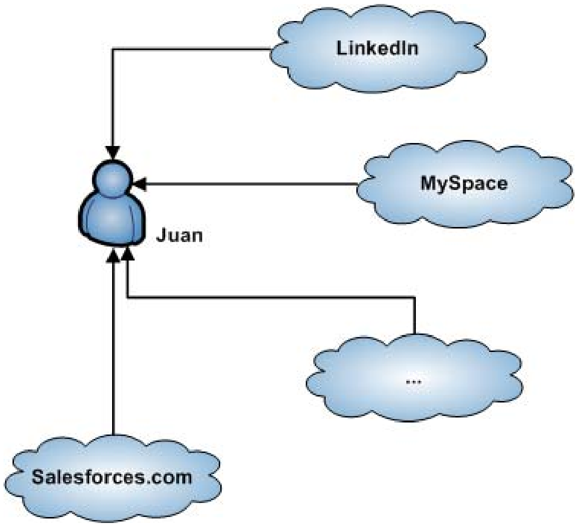


Fig. 2. Data Sources from Web 2.0 Applications

Nevertheless, in the past few years, there have been successful attempts of enriching tags with hierarchical relations [6] and the creation of faceted ontologies [7]. Furthermore, [8] describes the theory of formal classification, where labels are translated to a propositional concept language. Each node is associated to a normal formula that describes the content of the node, capturing g the knowledge that implicitly exists within simple classification hierarchies.

So, we can build an application that easily works across all the OpenSocial partners, and people who have an account in any social network supporting OpenSocial can use our solution for email ranking and filtering, taking advantage of the information in his/her social network.

4. Collaborative OSND-based ranking and filtering

The OpenSocial Network Dataset (OSND) is a lightweight ontology used for collaborative data filtering and rating in which we follow an integrated approach of combining three types of techniques for improving its construction from the tag sets gathered from the aforementioned Web 2.0 social networks such as Engage.com, Friendster, hi5, Hyves, imeem, LinkedIn, MySpace, Bebo, Ning, Oracle, orkut, Plaxo, Salesforce.com, Six Apart, Tianji, Viadeo, and XING.

The three techniques we are applying are as follows:

- Applying the Vector Space Model: The Vector Space Model [9] is an algebraic model used for information filtering, information retrieval, indexing and relevancy rankings. It represents natural language documents (or any objects, in general) in a formal manner through the use of vectors (of identifiers, such as, for example, index terms) in a multi-dimensional linear space. Documents are represented as vectors of index terms (keywords). The set of terms is a predefined collection of terms, for example the set of all unique words occurring in the document corpus. Relevancy rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as same kind of vector as the documents.
- Using Latent Semantic Analysis (LSA) [10] for analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA uses a term-document matrix which describes the occurrences of terms in documents. A typical example of the weighting of the elements of the matrix is the TF-IDF (Term Frequency–Inverse Document Frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up-weighted to reflect their relative importance.
- Validating the set of terms pertaining to the OSND with online lexical resources, such as Wordnet[1]. Dictionaries are generally considered as a valuable and reliable source containing information about the relationships among terms (e.g. synonyms). Also Wordnet can add conceptual meaning to the tags and there is an RDF transcript available.

Fundamentally, the coupling of the three techniques strongly founded on the Information Retrieval literature roots provide a two-pronged approach to retrieve and accurate OSND: selecting and extracting the most accurate tags from the pool of Web 2.0 applications user generated content and creating “metadata cloud” which encapsulates the subjective

[1] Wordnet: http://www.wordnet.com

meaning and intention the user conveyed through the tagging process. The OSND hence represents a valuable piece of knowledge which could be envisaged as a projection of the subjective mindset of the user.

5. Implementation

Primarily, to read email messages we need a client that has to be able to access to our email inbox. Most of the email's providers allow third party developers to get emails using POP3 and IMAP protocols.

The OSMail core (see Fig. 3) is a prototype email client based on OpenSocial that adds reputation ratings to the folder views of a message. This allows a user to see their reputation rating for each individual, and sort messages accordingly. This is, essentially, a message scoring system. While OSMail will give low scores to spam, it is unlike spam filters that focus on identifying bad messages. Another way to rate an email is comparing its content with the lightweight ontology, extracting the categories or the most relevant topics of itself.

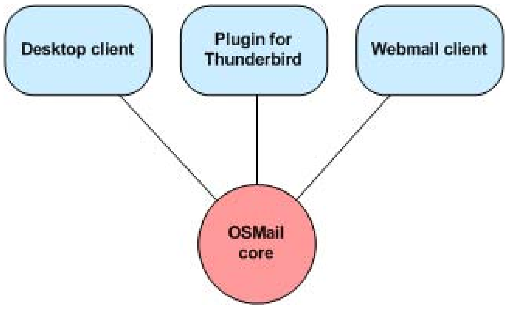


Fig. 3. OSMail prototype

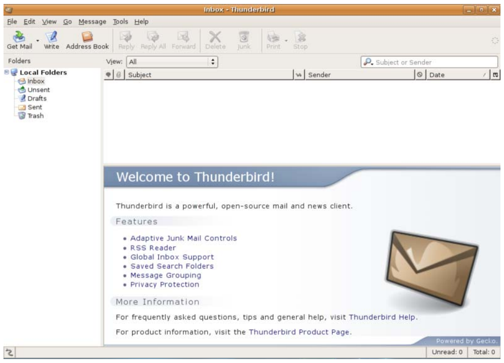
Therefore, its true benefit is that, in using the network, relevant and potentially important messages can be highlighted, even if the user does not know the sender.

Additionally, the solution can be oriented to the following approaches:

- A platform stand-alone mail application.
- A plug-in for Thunderbird.
- A webmail client.

Nevertheless, these three approaches (see Fig. 3) are only facades, in other words, user interfaces.

The first approach is a desktop application such as Mozilla Thunderbird (see Fig. 4) or Microsoft Outlook. We could develop another one, customized and optimized for the OSND. With a desktop application we can make your user interface look like almost anything we want and have total control over of the screen elements. Moreover, another advantage of this choice is that performance is generally quicker on a desktop because the screen is drawn only once on the desktop and only the data changes. This prevents a lot of screen data coming from the server to the client, which increases the time taken to display the data.



Welcome to Thunderbird!

Thunderbird is a powerful, open-source mail and news client.

Features

- Adaptive Junk Mail Controls
- RSS Reader
- Global Inbox Support
- Saved Search Folders
- Message Grouping
- Privacy Protection

More Information

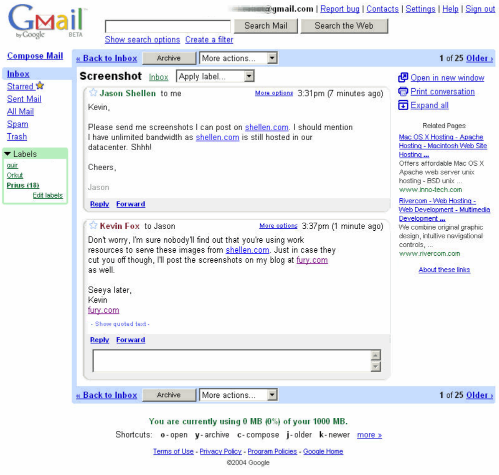
For frequently asked questions, tips and general help, visit Thunderbird Help.

For product information, visit the Thunderbird Product Page.

Fig. 4. The Thunderbird email client

The second approach is a plug-in for Thunderbird, maybe a good idea because users do not have to change their email client. A tremendous advantage of this alternative is the large community developing Thunderbird and lots of other plug-ins.

The third approach is a Webmail (see Fig. 5). It has several advantages, which include the ability to send and receive e-mail from anywhere using a single application: a web browser. This eliminates the need to set up an MTA/MRA/MDA/MUA chain.



[redacted]@gmail.com

Screenshot

Jason Shellen to me 3:31pm (7 minutes ago)

Kevin,

Please send me screenshots I can post on shellen.com. I should mention I have unlimited bandwidth as shellen.com is still hosted in our datacenter. Shhh!

Cheers,

Jason

Kevin Fox to Jason 3:37pm (1 minute ago)

Don't worry, I'm sure nobody'll find out that you're using work resources to serve these images from shellen.com. Just in case they cut you off though, I'll post the screenshots on my blog at fury.com as well.

Seeya later,
Kevin
fury.com

Related Pages

Mac OS X Hosting - Apache Hosting - Macintosh Web Site Hosting ...
Offers affordable Mac OS X Apache web server unix hosting - BSD unix ...
www.inno-tech.com

Rivercom - Web Hosting - Web Development - Multimedia Development ...
We combine original graphic design, intuitive navigational controls, ...
www.rivercom.com

About these links

You are currently using 0 MB (0%) of your 1000 MB.

Shortcuts: o - open y - archive c - compose j - older k - newer more »

Terms of Use - Privacy Policy - Program Policies - Google Home

©2004 Google

Fig. 5. The Gmail interface

Significant examples of email services which also provide the user a webmail interface are Hotmail, Gmail and Yahoo.

6. Related Work

[11] introduced the concept of a social network of cyberalter egos for distributing information among members of the network, and collaborative spam filtering. The network of cyberalter egos is constructed from the email addresses of the contacts in the users own address book. The program implements a percolation search algorithm to perform a scalable search of the network. The algorithm was tested on a real email dataset, and achieved a spam detection rate of almost 100% with a minimized false positive rate.

[13] present an automated graph theoretic method in order to extract the trusted social network of an email user based on the email addresses of senders and recipients in the email headers. Subsequently the information provided by the user's social network is effectively used to classify emails as spam or non-spam. The authors claim an effectiveness rate of 100% with no false negatives, with 53% of all emails being classified as spam or non-spam.

[14] present an algorithm to provide a reputation rating for each of the contacts in a user's social network. Based on the ratings, a reputation network with reputation values between each of the individuals in the user's network can be constructed. Using reputation scores for the sender of each message which arrives in a user's inbox, emails are filtered according to priority.

The uniqueness of our approach compared to previous work lies in the fact that we use the OpenSocial platform to extract the relationships in a user's social network, and therefore the algorithm can attach importance weights based on a user's contacts across all social sites which support open social. This is a novel approach compared to previous spam filtering techniques, which constructed the user's network based on message exchange in a user's email account.

7. Conclusions and Future Work

In this paper we have presented a novel way of ranking and filtering email based on a user's social network using the Google OpenSocial API. Our future work intends to evaluate the accuracy of the algorithm using established metrics such as precision, recall and the F-measure. A further objective is to test the efficiency of the algorithm on a number of use cases, and thus diverse email data sets.

An additional problem which occurs when mining email, which was investigated by [15], is the use of various email aliases by a single user. Our future work tends to investigate the effect of this activity in email data sets.

In the future we intend to use the algorithm as a plug-in for email filtering and ranking of popular email clients such as Thunderbird and Outlook.

A preliminary conclusion is that we expect that the algorithm will improve email classification efficiency from a user perspective, as it is based on the Social Network of the user. We hope to prove this in future evaluations of the classification performance of the algorithm.

8. Acknowledgements

This work is supported by the Spanish Ministry of Industry, Tourism, and Commerce under the GODO project (FIT-340000-2007-134), under the PIBES project of the Spanish Committee of Education & Science (TEC2006-12365-C02-01) and under the MID-CBR project of the Spanish Committee of Education & Science (TIN2006-15140-C03-02).

9. References

[1] Fensel, D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag. 2002.

[2] Ankolekar, A., Krötzsch, M., Tran, T., and Vrandecic, D. 2007. The two cultures: mashing up web 2.0 and the semantic web. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM Press, New York, NY, 825-834.

[3] University of California Police Department. You’ve Got Spam: How to Avoid Unwanted Email.

[4] OpenSocial – Google Code official web site. http://code.google.com/apis/opensocial/

[5] Shadbolt, N. Hall, W. Berners-Lee, T. The Semantic Web Revisited. IEEE Intelligent Systems. 2006.

[6] Schmitz, P. Inducing Ontology from Flickr Tags. Collaborative Web Tagging Workshop. Proceedings of the 15th WWW Conference. 2006.

[7] Heyman, P. Garcia-Molina, H. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report Stanford University. 2006.

[8] Giunchiglia, F. Marchese, M. Zaihrayeu, I. Towards a Theory of Formal Classification. Proceedings of the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications. Pittsburgh, Pennsylvania. 2005.

[9] Salton, G. Wong, A. and Yang, C. S. A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, nr. 11, pp 613–620. 1975.

[10] Deerwester, S. Dumais, Furnas, G. W. Landauer, T. K. Harshman, R. "Indexing by Latent Semantic Analysis". Journal of the Society for Information Science 41, Issue 6. pp 391-407. 1990.

[11] Joseph S. Kong, Behnam A. Rezaei, Nima Sarshar, and Vwani P. Roychowdhury. Let Your CyberAlter Ego Share Information and Manage Spam. 2005.

[12] Gomez, J.M. Colomo, R. Alor-Hernandez, G. Posada-Gomez, R. Garcia, A. Search in the Eye of the Beholder: Using the Personal Social Dataset and Ontology-guided Input to Improve Web Search Efficiency. Proceedings of the 5th IEEE Latin-American Web Conference (LA-WEB07). Santiago de Chile, Chile. October, 31- November, 2nd. 2007.

[13] Boykin, P., Roychowdhury, V. Personal Email Networks: An Effective Anti-Spam Tool. IEEE Computer, Vol. 38, No. 4, pages 61-68. 2005.

[14] Golbeck, J. and Hendler, J. Reputation Network Analysis for Email Filtering. Proceedings of Conference on Email and Anti-Spam. Mountain View, California, USA. 2004.

[15] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, Anand Swaminathan. Mining Email Social Networks. MSR’06, May 22–23, 2006, Shanghai, China.

[16] Golbeck, J. and Hendler, J. Reputation Network Analysis for Email Filtering. Proceedings of Conference on Email and Anti-Spam. Mountain View, California, USA. 2004.

[17] Ankolekar, A., Krötzsch, M., Tran, T., and Vrandecic, D. 2007. The two cultures: mashing up web 2.0 and the semantic web. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM Press, New York, NY, 825-834.

[18] Gomez, J.M. Colomo, R. Ruiz, B Garcia, A. ProLink: A Semantics-based Social Network for Software Project. In International Journal of Information Technology and Management. Special issue: Work Change in the Era of ICTs. 2007.