Swimming against the *Streamz*: Search and Analytics over the Enterprise Activity Stream

Ido Guy, Tal Steier, Maya Barnea, Inbal Ronen, Tal Daniel
IBM Research Lab
Haifa 31905, Israel
{ido, talst, mayab, inbal, taldan}@il.ibm.com

ABSTRACT

Activity streams have become prevalent on the web and are starting to emerge in enterprises. In this work, we present *Streamz*, a novel application that uses a faceted search approach to provide employees with advanced capabilities of search, navigation, attention management, and other types of analytics on top of an enterprise activity stream. We provide a detailed description of the *Streamz* tool as well as usage analysis based on user interface logs and interviews of active users.

Categories and Subject Descriptors: H.5.3 [Group and Organizational Interfaces]: Computer-supported cooperative work

General Terms: Design, Experimentation, Human Factors.

Keywords: Activity streams, enterprise, social media. .

1. INTRODUCTION

The recent evolution of the web, often referred to as the *real-time web*, is characterized by highly intensive streams of updates and news. Leading social media sites, such as Facebook, Twitter, LinkedIn, Myspace, and Google+, publish activity streams that include millions activities per day, generated by millions of users who write status updates, share links and photos, join groups, comment, and "like" others' activities.

Following their prosperity on the web, social media applications, such as wikis, forums, social bookmarking, file sharing, or blogging systems, have also emerged within the enterprise, enabling employees to share and interact behind the firewall. As part of the proliferation of enterprise social media, activity streams that syndicate employees' activities across the organization's social media have also started to emerge [5,7,8].

The emergence of activity streams within the enterprise poses a great opportunity in terms of search and analytics. This unique medium of highly intensive activities, concise in text and metadata, allows employees to stay tuned with recent updates and to discover new developments that relate to their areas of interest. Moreover, these streams can help increase awareness of organizational projects and processes, and expose recent trends and opinions. In a global enterprise, with many distributed teams working in different locations and substantial time-zone differences, the value of this kind of social awareness can be especially high.

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10...$15.00.

In this work, we present *Streamz*, an application we developed to provide search and analysis of enterprise activity streams. We implemented *Streamz* on top of our organization's activity stream, which syndicates activities across a wide variety of enterprise social media applications, including blogs, microblogs, wikis, files, bookmarks, and forums. All of these applications have been deployed in our organization for over three years and enjoy a large user base and frequent activity. Within our company, an activity, defined as the basic unit of the stream, occurs at a frequency of under four seconds during working hours.

Streamz provides real-time search [3] on top of the activity stream, allowing users to search activities seconds after they have been posted, typically sorting the results reverse-chronologically. Faceted navigation enables users to apply multiple filters to explore the stream. Advanced analytics, such as personalization via a user interest profile, topic extraction, sentiment detection, and activity grouping, provides further means for gaining insights from the stream.

We describe *Streamz* in detail, including its users interface, analytics components, and backend. We also analyze the way the different features were used through inspection of *Streamz'* user interface logs over a period of two months, where 239 authenticated users performed a total of 4,135 actions. To complement the analysis, we interviewed 10 of these users.

2. RELATED WORK

A few recent studies examined activity streams in the enterprise, focusing on personalization techniques. Freyne et al. [7] proposed a method for narrowing the stream of the SocialBlue enterprise social network site based on person and action relevance inferred from users' browsing behavior. Daly et al. [5] suggested viewing the activity stream through "social lenses", based on user-defined collections of people and entities. Guy et al. [8] studied personalization of the stream based on a user model that includes people, terms, and places of interest. Our work does not focus on evaluating methods for stream personalization, but rather provides a broad application overview and analysis to better understand the potential value of search and analytics over the enterprise stream.

This work is not the first to propose the use of faceted search to navigate aggregated feeds. The Eddi tool [2] represented topics as facets and suggested an alternate Twitter interface that allowed topic-based browsing. Visual Backchannel [6] applied a faceted search approach on microblog conversations to show related people, images, and topics. Perhaps the most relevant work is by Hong et al. [11], who introduced FeedWinnoer, an enhanced feed aggregator that allows knowledge workers to filter feed items by topic, person, source, and time. Only an initial evaluation was provided, based on interviewing 15 employees. In this work, we explore an activity stream of social media behind the firewall,

rather than external feeds of websites or microblogging services. The stream originates from a wide variety of enterprise social media applications, such as blogs, wikis, and bookmarks. Several types of analytics, such as activity grouping, sentiment analysis, and the use of personalized profiles to filter the stream, have not been applied by the tools above.

3. THE *STREAMZ* APPLICATION

We built *Streamz* on top of the activity stream of IBM Connections (IC) [12]. IC is a social media application suite for the enterprise, which includes eight types of applications, all of which have been deployed in our organization for over three years: blogs, bookmarks, files, forums, microblogs, profiles, tasks, and wikis. IC publishes an activity stream of all public actions occurring in its applications, e.g., creating, editing, commenting, or "liking" of a blog post or a wiki page; creating or replying on a forum thread; or connecting, following, or tagging another person on profiles.

3.1 Design Goals

Streamz was designed to help users consume an activity stream, with the following key goals:

- Attention management – help the user surface the "interesting" activities out of the stream.
- Search – enable the user to find an activity based on a few details he remembers, such as author or keywords.
- Navigation – enable the user to easily move from one list of activities to another. For example, the user sees that someone has edited a wiki of interest and may wish to see all the wikis that person has edited.
- Big picture – enable the user to look at an aggregated view of the activity stream or a subset of it to gain broad insights. For example, a user may wish to look at the subset of the stream dealing with a particular project and understand the sentiments regarding it.

Moreover, since the activity stream is temporal by nature, it was crucial to support the freshness of data being displayed.

3.2 User Interface

Opening the *Streamz* application for the first time displays the most recent activities across the entire organization. The user is prompted to authenticate with his intranet password to enjoy and control personalization features. A persistent cookie is used to identify returning users who previously authenticated. By default, the activities are auto-refreshed every ten seconds, but the user can choose to disable this. A search box at the top of the page allows the user to enter a query, narrowing the stream to activities relevant to that query. The search supports full free-text syntax, including Boolean operators such as OR, AND, NOT, and phrase queries using double quotes. Activity results are paginated, with ten activities per page. The user can navigate through the result pages via links at the bottom of each page.

Figure 1 illustrates the main user interface of *Streamz*. The user can filter the activity stream in various ways, but ultimately one user interface is used to present any subset of the stream. The UI consists of three main parts. Its main component, the stream (on left), displays the activities in the stream based on the filtering criteria currently applied. The facets, on the right, summarize various aspects of the stream, such as leading topics and active people. The upper section displays the current user profile for authenticated users and may be hidden. Below, we describe each of the three UI components in detail.

Stream. Activities are displayed in reverse-chronological order. If the user has entered search terms, they are highlighted in light yellow. Each activity includes a picture of its author and an icon indicating the originating IC application, which links to the activity's URL. The text of the activity may include a description and an excerpt of the content. Each underlined entity within the activity description is a link to its corresponding IC page. Below the text of the activity is an indication of its freshness, such as "4 hours ago" or "yesterday". Next to that, a bar indicating the sentiment level may appear, if the activity has been identified as either positive (green) or negative (red). Further below is a "more from" line that allows the user to re-filter the stream to any of the entities in the current activity, with a number indicating how many activities relate to each entity.

Facets. In the *faceted search* approach, especially common on e-commerce websites, the user can explore and navigate the search results based on various predefined categories, known as facets [1]. The facets allow the user to both get a better summary of the results by presenting the most common values in each facet, usually with some indication for the number of corresponding search results, and to drill down (refine) the results by choosing a certain value of a certain category. In *Streamz*, we defined five types of facets on top of the activity stream data: topics, sentiment, source, people, and timestamp. The timestamp facet is not visually presented, but enables temporal visualizations since it allows categorizing the results by time range. The other four facets are always presented to the right of the activities, summarize the current stream, and allow refining it by a specific value. In Figure 1, the user has originally searched for "John Smith" and then further refined the results to "Social Analytics", leading to a display of activities that match the query "John Smith AND Social Analytics".

The topic facet presents the most common topics in the stream in two ways. The upper graph shows a time-based visualization of the top three topics in the stream over the past week. The word cloud below it presents the top 20 topics, with the size of each proportional to its facet value score. Clicking on a topic in either display refines the results to only those relevant to the selected topic. Our topic extraction method is explained in the Analytics section.

The sentiment facet, presented as a bar, shows the partition of the activities in the stream into three sentiment classes: negative, neutral, and positive. The sentiment class of an activity is determined by its sentiment score, calculated as described in the Analytics section. The bar gives an indication of the overall sentiment regarding the current stream and allows the user to refine the query to include only results with a specific class of sentiment.

The source facet shows the partition of the activity results according to the originating application. For each source, it indicates the portion of the stream that originates from it, and the user can refine the stream to a specific source. The person facet indicates the most active people on the stream and enables refining the results to a specific individual.

Profile. The user profile provides the means for personalization and aims to filter the stream to a mix of activities according to the user's interests. Contrary to leading stream applications such as Twitter and Facebook, which base their personalization on the set

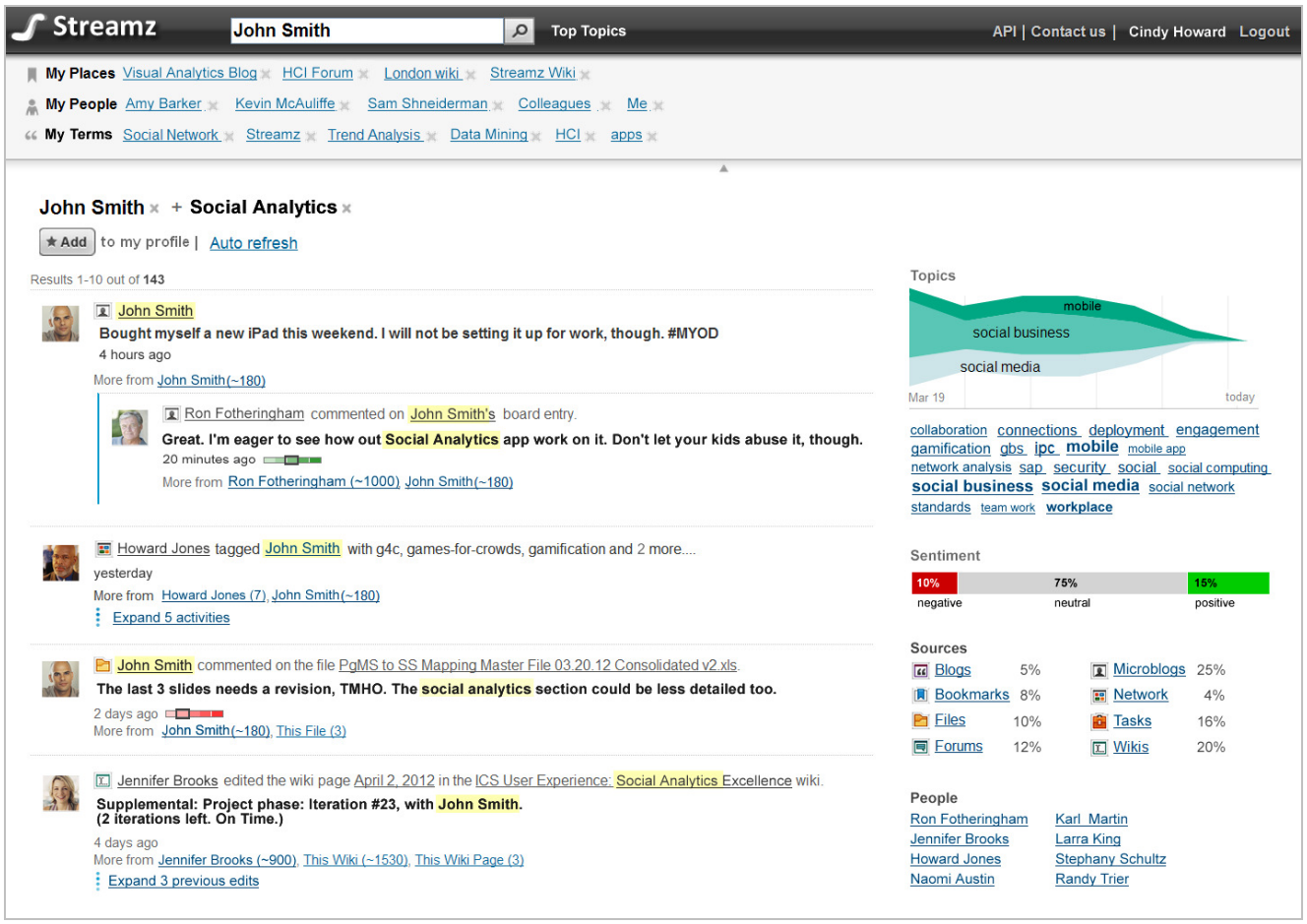

Figure 1. *Streamz* main user interface.

of people the user follows or has as friends, *Streamz* personalization is based on three types of profile items: people, terms, and places. Places are entities for which multiple activities can occur, such as blogs and blog posts, files, forums and forum threads, microblog threads, tasks, and wikis and wiki pages. This approach is motivated by the results of a previous study that examined the personalization of the activity stream [8] and found that each of the three items is useful for personalization and brings to the table a different mix of desired qualities.

When the user authenticates, his profile is presented at the top of the screen, as shown in Figure 1. Items are added to the profile by saving search queries. When the user issues a query not yet in his profile, an "Add to my profile" button appears, enabling the user to change the descriptive name of the query and save it to his profile as a person, term, or place. When the profile is blank, the user is prompted to add items to it by recommendation of related people and terms [10]. Recommendation of places is currently not supported. For authenticated users with a non-blank profile, the default view of the stream is personalized to only include activities that relate to at least one of their profile items.

The user can click on any profile item to initiate a query and filter the stream to activities that relate to that particular profile item. Overall, *Streamz* provides four ways to search the activity stream: a manual search query, via a profile item, via "more from", and via a facet. The first three initiate a new query, while the latter refines the current results.

3.3 Analytics

Topic extraction. For extracting the topics for a given activity, we used the Kullback-Leibler (*KL*) measure, which identifies a set of terms that maximizes the divergence between the language model of the activity and the language model of the entire stream. In addition, we applied a tag-boost (*TB*), which promotes keywords that are likely to appear as tags, based on a given well-tagged folksonomy. We used the IC bookmark application folksonomy for this purpose. The *KL+TB* method has been found to be highly effective for term extraction in non-tagged domains [4] and enables the extraction of both single and multiple-word topics. The weighted list of an activity's related topics was generated by applying *KL+TB* on its text, after filtering out people's names and reserved keywords and stemming.

Sentiment Detection. We used the *Apache UIMA* framework for content analytics (http://uima.apache.org) to calculate a sentiment score for each activity in the stream. Our *UIMA* annotator uses a language processing engine to identify sentiment segments in the text, i.e., segments that are associated with a non-neutral sentiment. For example, the segment "not good" would be associated by the engine with negative sentiment. The overall sentiment score of the activity is calculated by the average score of the sentiment segments it contains. While sentiment analysis evaluation is beyond the scope of this work, our initial tests showed a level of accuracy of roughly 70%.

Activity Grouping. The activity stream often contains groups of activities that belong together, since they are similar to one another or jointly constitute a more complex structure. We implemented three types of activity groupings in *Streamz*: (1) The *Thread* grouping groups together a message and all replies. It has been implemented for microblogs only. The thread is presented in chronological order, with the first message appearing at the top and up to two indented replies, selected based on these priority criteria (listed in order): (a) matching the current search query and (b) being more recent. In Figure 1, the first activity is a microblog thread, with one message and one reply. (2) The *Duplicate* grouping groups together identical activities that have different timestamps. We implemented duplicate grouping for wikis and blogs in the case of multiple consecutive edits of the same page by the same user. In Figure 1, the last activity is a duplicate of four wiki page edits. The duplicate is presented by its most recent activity. (3) The *Compound* grouping groups activities that have the same author and action, but different objects. We implemented compound grouping solely for person tagging in case a user has tagged the same person with multiple tags. In such cases, the activity's text would change to reflect the summarization of all tagging activities, e.g., "p1 tagged p2 with t1, t2, and t3". In Figure 1, five tagging activities are grouped together. For all grouping types, the user can expand to all single activities (and collapse back to the original display) by clicking a link.

3.4 Backend

Streamz is built on top of Lucene [14], a popular open-source search engine library. Lucene is used for indexing, where the indexed documents are activities retrieved from the IC stream. Grouped activities are treated as single activities, with a timestamp corresponding to the most recent activity in the group, and include a link to the list of single activities that compose them. Lucene's faceted search capabilities are used to support the five types of facets described above. Lucene also enables to efficiently calculate the "more from" counts using the document frequency attribute.

Activities are retrieved by continuously polling (every ten seconds) the IC activity stream for the most recent activities. Before indexing, analytics is applied on each activity for topic extraction, sentiment level detection, and potential grouping with previous activities. All user searches in the *Streamz* UI are translated into queries and executed over the index. The personalized view is calculated by executing a Boolean OR query between all of the user's profile items.

Streamz architecture partitions the index into several small indices that reside in memory and hold the more recent activities and a few larger indices that reside on the file system and hold the older data. Only one "active" index exists to hold the most recent activities of up to the last 24 hours. We use Lucene's real-time indexing capabilities to refresh this small memory-resident index in a matter of a few milliseconds, so that parallel search requests can be seamlessly served. Traversing the indices is performed from the freshest to the oldest and is stopped when the results span at least a week (for robust facet calculation and enablement of the topic graph facet) *and* when enough results exist to fill the current page requested by the user. This approach often saves the need to traverse the entire index, since the two completion conditions are frequently satisfied by traversing the first few indices. Overall, the backend architecture enables activity appearance on the *Streamz* UI up to 15 seconds after they have been posted.

4. USAGE ANALYSIS

Our usage analysis is primarily based on the user interface logs from *Streamz*, which document every user action along with a timestamp and the user's ID if they are authenticated. We analyzed the logs recorded from January 15, 2012 to March 14, 2012. During this period, 239 distinct authenticated users used *Streamz*. We also conducted 30-minute phone interviews with 10 of them to gain in-depth insight of their use of *Streamz*.

When asked about the most compelling features of *Streamz*, our interviewees noted the ability to customize the stream using saved queries and other filters, the facets that summarize the stream and allow narrowing the current context, the diversity of sources, the search result snippets and search term highlighting, and the auto-refresh. A few mentioned that they keep a browser tab open with *Streamz*, so they can continuously monitor updates of interest.

Figure 2 details the main action types logged within *Streamz* and their distribution of occurrence over the time period (percentage of the total number of actions). The average number of actions per user was 17.3 (stdev: 28.29, median: 7, max: 224). Searching using the main search box was the most popular action (over 34% of the actions), followed by clicking on an external link and clicking on a result page number. Facet clicking and profile editing were also common. At the bottom of the list are clicks on profile items, "more from" links, and expanding or collapsing grouped activities. In the remainder of this section, we analyze some of the more common action types in more detail.

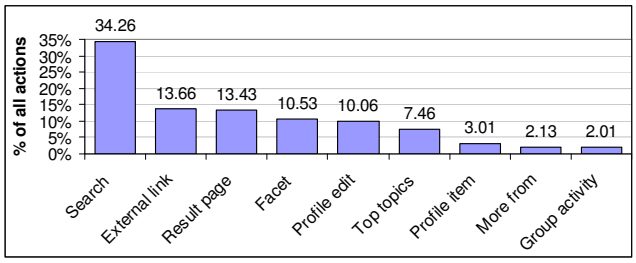

Figure 2. Action distribution by type.

External Links. We refer to all links within an activity that lead to IC as external links. Previous studies on traditional web search have found that 50% to 87% of the searches yield a click on at least one result [9]. Looking at Figure 2, we can see that in *Streamz*, the ratio of clicks on external links is substantially lower, indicating that *Streamz* provides enough information to often spare users the need to go to IC. Several interviewees pointed this out, e.g., "*Usually I see all the information I need in Streamz*" and "*Since I read the news sequentially, clicking on a link would hurt my flow*". Clicks on wiki activities were the most common, in a similar proportion to their frequency in the general stream. Microblogs were clicked much more often than their frequency in the stream (20.85 of all clicks for only 9.45% of the stream). One interviewee explained, "*I often go to a person's board to see the broader context […] sometimes the conversation relates to a previous thread on the board or arouses my curiosity to read previous messages.*" On the other hand, in spite of their high frequency, profile-related activities (follow, tag, connect) were rarely clicked, apparently since they have no content.

Result Pages. Almost half of the page navigations went to the second page of the results, and 18.7% to the third page. In general, viewing additional result pages was more common than in content search, where it is well known that users rarely access pages beyond the first. It is likely that in content search users look

for a specific result, while in *Streamz* they often scroll through a sequence of fresh activities related to their query.

Facets. Figure 3 shows the distribution of clicks by the different facet types. The source facet was the most commonly used, followed by the topic facets (the cloud and the graph). Hong et al. [11] found that topic facet was mentioned as the most useful (9/15 interviewees), while source facet only came second (4/15). Our findings suggest that in practice, source facet was the most useful, even though we suggested two types of topic facets. Person facet was used relatively infrequently, possibly due to its location at the bottom of the facet list. For the source facet, files were the most popular source clicked. Some interviewees mentioned that they sometimes looked for a particular deck or a photo that was shared through the files application. Microblogs were the second most popular source, while profiles were the least popular.

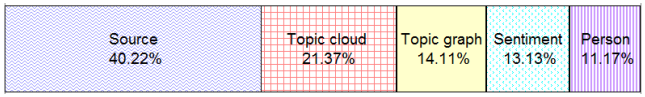

Figure 3. Facet click distribution.

The sentiment facet was mentioned by several interviewees as useful. One said "*I like the sentiment facet since it gives me a sense of people's opinion on the subject*" and another mentioned "*I use it to evaluate the type of comments I get about my projects*." Of the clicks on the sentiment facet, 29.8% were for positive, 57.4% for negative, and 12.8% for neutral. Apparently, the negative filter arouses more curiosity than the positive. One interviewee said "*I use the negative filter to track issues or problems with my project [...] I use the positive filter to see nice comments about my project, which I can use when I pitch it to executives or customers*." Another interviewee commented "*As a business consultant I need both sides of the story to build an argument. [It] would be nice to have a view of the positive versus the negative side by side, like in Amazon reviews."*

"More from". The majority of the "more from" clicks—63.6%—was performed on people, serving as another means to explore the stream of a specific individual. The rest were performed on places, distributed rather evenly across forums, files, blogs, and wikis.

Grouped activities. The vast majority of expansions and collapses of grouped activities referred to threaded activities (microblogs)—79%. Only 9.9% and 11.1% referred to duplicate wiki or blog edits and "compound" person tagging, respectively, indicating these usually remain collapsed.

Profile items. 76.1% of *Streamz* authenticated users had at least one item in their profile. On average, each of these users had 7.21 terms (stdev: 5.92, median: 7, max: 34), 3.75 people (stdev: 5.59, median: 0, max: 42), and 0.2 places (stdev: 1.2, median: 0, max: 15) in his profile. The low occurrence of places is likely since we only provided recommendations of people and terms. This might have also been affected by the fact that the only way to query for a particular place is through the "more from" link. Since places have been shown to produce a high portion of interesting activities [8], further measures need to be taken to promote place addition to user profiles. One interviewee suggested: "*It would worth to recommend wikis and forums in which I've been active recently […] so I can add to my profile, even for a limited time."*

Many users—40.9% of those who had a profile—included themselves as part of it. From our interviews, we found this to have been motivated by two main uses: "*I click on myself to search my own activity, for example to find something I remember I shared or liked*," said one interviewee, while another explained: "*I'd like to see whenever someone mentions my name.*" The most popular profile item was "web 2.0" (16 user profiles), followed by two internal product names (13 and 12 profiles, respectively), an internal group name (11), "social business" (10), and "android" (10). Out of all profile item clicks, 66.34% were on terms, 29.27% on people, and 4.39% on places.

Profile edits. 68.89% of the profile edits referred to the addition of new profile items, while 31.11% referred to the removal of items. The high ratio of profile edit actions and the relatively high percentage of item removals indicate the dynamism of the profile and users' need to continuously update it. A similar phenomenon has been observed for the list of a user's followees on Twitter [13]. One interviewee explained: "*when I add a topic or a person and realize it is too noisy, I will remove them*." Another interviewee noted, "*When I start a new project, I update my profile with new terms and people that relate to it*."

5. CONCLUSION

This paper provides a first glimpse into the use of an activity stream in the enterprise. With younger individuals joining the workforce, accustomed to using tweets and news feeds as their main means for interaction, an integrated activity stream, equipped with search and analytics capabilities, is likely to play a central role in the future shape of workplace collaboration.

6. REFERENCES

[1] Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek, S., Sheinwald, D., Shekita, E., Sznajder, B., & Yogev, S. 2008. Beyond basic faceted search. *Proc. WSDM '08*, 33-44.

[2] Bernstein, M.S., Suh, B., Hong, L., Chen, J., Kairam, S., & Chi, E.H. 2010. Eddi: interactive topic-based browsing of social status streams. *Proc. UIST '10*, 303-312.

[3] Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., & Lin, J. Earlybird: real-time search at Twitter. *Proc. ICDE '12*.

[4] Carmel, D., Uziel, E., Guy I., Mass, Y., & Roitman, H. 2011. Folksonomy-based term extraction for word cloud generation. *Proc. CIKM '11*, 2437-2440.

[5] Daly, E.M., Muller, M., Millen, D.R., & Gou, L. 2011. Social lens: personalization around user defined collections for filtering enterprise message streams. *Proc. ICWSM '11*.

[6] Dork, M., Gruen D., Williamson, C., & Carpendale, S. 2010. A Visual Backchannel for Large-Scale Events. *IEEE Trans. Vis. and Comp. Graphics* 16, 6 (Nov. 2010), 1129-1138.

[7] Freyne, J., Berkovsky, S., Daly, E.M., & Geyer, W. 2010. Social networking feeds: recommending items of interest. *Proc. RecSys '10*, 277-280.

[8] Guy, I., Ronen, I., & Raviv, A. 2011. Personalized activity streams: sifting through the "river of news". *Proc. RecSys '11*, 181-188.

[9] Guy I., Ur, S., Ronen I., Weber, S., & Oral, T. 2012. Best faces forward: a large-scale study of people search in the enterprise. *Proc. CHI '12*, 1775-1784.

[10] Guy, I., Zwerdling, N., Ronen, I., Carmel, D. & Uziel, E. 2010. Social media recommendation based on people and tags. *Proc. SIGIR '10*, 194-201.

[11] Hong, L., Convertino, G., Suh, B., Chi, E.H., & Kairam, S. 2010. FeedWinnower: layering structures over collections of information streams. *Proc. CHI '10*, 947-950.

[12] IBM Connections – Social Software for Business: http://www.ibm.com/software/lotus/products/connections

[13] Kwak, H., Chun, H., & Moon, S. 2011. Fragile online relationship: a first look at unfollow dynamics in twitter. *Proc. CHI'11, 1091-1100.*

[14] McCandless, M., Hatcher, E., & Gospodneti, O. 2010. Lucene in action, 2nd edition. *Manning Publications Co.*