IBM Research Report

Contextual Analysis of User Interests in Social Media Sites – An Exploration with Micro-blogs

Nilanjan Banerjee[1], Dipanjan Chakraborty[1], Koustuv Dasgupta[1], Anupam Joshi[1,#], Sameer Madan[2,*], Sumit Mittal[1], Seema Nagar[1], Angshu Rai[3,*]

[1] IBM Research - India
4, Block - C, Institutional Area, Vasant Kunj
New Delhi - 110070, India.
{nilanjba, cdipanjan, kdasgupta, anupam.joshi, sumittal, senagar1}@in.ibm.com

[2] Indian Institute of Technology Delhi
Hauz Khas
New Delhi - 110016, India.
sameer27.in@gmail.com

[3] Indian Institute of Technology Madras
Dept. of Computer Science and Engineering
Chennai, India.
angshu@cse.iitm.ac.in

[#] Work done at IBM Research while the author was on leave from University of Maryland, Baltimore County, USA

[*] Work done while being at IBM Research - India, New Delhi

IBM Research Division
Almaden - Austin - Beijing - Delhi - Haifa - T.J. Watson - Tokyo - Zurich

Contextual Analysis of User Interests in Social Media Sites – An Exploration with Micro-blogs

ABSTRACT

Since their inception, social media sites have evolved from instant messaging and online networking to diverse, multi-faceted entities that encompass the entire personality of an individual. Recent advances in technology around mobile–based access to social networking platforms and facilities to update status information in real–time (e.g. in Facebook) have further allowed an individual's online presence to be as ephemeral and dynamic in nature as her very thoughts and interests. In this context, micro-blogging has been widely adopted by users as an effective means to capture and disseminate their thoughts and actions to a larger audience on a daily basis. Interestingly, the daily chatters of a user obtained from her micro–blogs offer a unique information source to analyze and interpret her *context* in real-time – i.e. interests, intentions, and activities. Rich contextual information about users allow social networking players to develop value–added applications and associated business models to monetize the same.

In this paper, we gather data from the public timeline of Twitter (one of the most popular micro-blogging sites) spanning across ten worldwide cities over a period of four weeks. We use this dataset to (a) explore how users express interests in real–time through micro–blogs, and (b) understand how unstructured text mining techniques can be applied to interpret the real-time context of a user based on her *tweets*. Our findings provide evidence that social media sites like Twitter constitute a promising source for extracting user context that can be exploited by a multitude of social networking applications.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Experimentation

Keywords

Social Networks, Micro–blogs, Text Mining

1. INTRODUCTION

Social media is part of a growing trend in the web, where a majority of the new content being added is user generated. Social media includes a variety of sources where people share their thoughts (e.g. blogs, microblogs), multimedia (e.g. YouTube, Flickr), and even their personal information (e.g. Facebook, Orkut). Social networking sites allow people to not only mimic their real life networks online, but cross geo-spatial and social boundaries to create networks that would not be otherwise possible.

Like with any other technology, what attracts people initially is not what keeps them vested in the long run. This is especially true for one driven by the community itself. Hence, there would be a tipping point where social networking users get tired, lazy or lose the initial excitement and churn out. While it is debatable when this point would be reached, social media sites are already looking at strategic initiatives to sustain and monetize their networks. We believe that this would largely involve an image make–over of social networks from its broad–based usage (e.g. writing on the wall, browsing, viewing photos and "poking" each other) to a more *value–centric* one that could possibly include search, advertising, commercial transactions, and other productivity–focused behavior.

To provide such value-adds, most agree that it is inevitable for online social media sites to open up their services for integration with third parties [3]. With increasing linkages between mobile and Internet services, Telecom operators are simultaneously looking at ways to harness the potential of social networking. The efforts are not limited to bringing social sites to a mobile format. Operators have a wealth of content associated with their network, along with core network enablers, like converged communications, location, and presence. Consequently, "contextual communications" is emerging as one of the key features of social networks, i.e. where consumers are actively using end device (mobile, PDA, laptop) and a plethora of communication modalities (SMS, Google Talk, Facebook status) – updating their status, availability, mood and even current interests to drive enhanced levels of communication. Rich *context* of a user is thus not limited to her location and/or availability, but captures attributes that directly extend a user's personality. Further, one can envision this rich presence being catered to social networking communities – thereby breeding a new genre of real–time collaboration. Such collaboration

.

(a movie outing or a business seminar) not only leads to value–addition, but also opens up untapped revenue streams for the social networking site (e.g. each time a movie is scheduled, revenue is collected from the theater).

The combination of ubiquitous connectivity with online social interactions is leading to a plethora of new applications in "real-time" social networks or "geo" social networks. New applications such as Ning[4], Plazes[5] and X [12] allow social interactions that are a function of the users spatio-temporal context. For instance, X allows users to connect with others in their social network who are geographically close, and available for some activity of mutual interest (e.g. *which of my friends are interested in watching the new Star Trek movie this evening in the theater downtown?*). Whenever the context is limited to spatio-temporal elements, these are (relatively) easily and precisely obtained (e.g. assisted GPS type devices or E911 type systems). However, as the example shows, some elements of context involve personal preferences and plans of the requester and those of others in his/her social network. Interestingly, we can leverage elements of social media to automatically infer a user's interests, as well as her short term plans. In this context, we focus on the recent micro–blogging phenomenon as a potentially rich source of user context. Perhaps the most commonly used micro–blogging platform is Twitter[6], and its popularity is attested to by the fact that twittering is now commonly used as a verb! While many refer to microblogging as a manifestation of the increasing narcissism of modern society, the tweets of an individual provide a window into their own perspective of what they are thinking and doing. As such, an analysis of the microblogs can potentially let us infer what interests a person, and even what they might be doing or thinking of doing. Clearly, these two elements form an important part of the context of the user in real time geo-social networks.

Challenges and Contributions: The ability to obtain short terms plans of a user is clearly a novel capability that analysing a user's microblogs provide. The ability to infer interests from text however has been done before – e.g. from analysing the users web page, or even their blogs. What makes the problem challenging and different is that tweets tend to be stream of consciousness fragments[1]. They lack the structure of a cogent argument or description, even to the rather limited extent it is found in blogs. Moreover, even more so than blogs, tweets (even the public ones) are directed at friends and family who share a common frame of reference. Many things can therefore be said or implied without being explicitly spelled out – including place names and times. Consider a tweet which says "Going to dinner this evening at the usual Chinese place". People who know the tweeter are likely to know what this place is and also when the user will go out. This, of course, is not evident from the text.

To support our contention that microblogs can be analysed to infer interests and activities of a user, we gathered data from the public timeline of twitter for ten cities in different parts of the world. For each city, the tweets of the top thousand tweeters were gathered for a period of four weeks. In this paper, we report the results of our analyses of these tweets using unstructured text mining techniques. We show that these tweets let us infer not only the interests of the user, but also allow contextual insights into their planned activities. We gain such insights by mining latent associations present in "free-text" tweets that indicate real-timeliness of user interests. In doing so, we employ a staged approach – first, we discover representative keywords that describe interests along multiple dimensions. Second, we employ several statistical techniques to discover associations between these dimensions. Finally, we analyze patterns that are present in select clusters containing real-time interest tweets. We believe that the findings presented herein are a first-of–a-kind, in terms of uncovering contextual information through micro–blogs analysis, and a right step in the direction of truly ubiquitous social networking – that seamlessly integrates with the social context in our communications, our interests, and our activities.

[1]http://www.fastcompany.com/magazine/127/scobleizer-stream-of-consciousness.html

Table 1: Twitter Data Statistics

Cities	No. of Tweets (in million)	No. of Users
New York	3.65	109548
Los Angeles	2.74	61445
London	2.42	59810
San Francisco	2.09	38691
Chicago	1.85	32908
Atlanta	1.72	32933
Toronto	1.64	34125
Boston	1.58	32723
Seattle	1.45	30230
Austin	1.14	22489

2. DATA SET

Twitter [6] is a free social-networking service that enables friends, family and colleagues to stay connected through sending and receiving updates known as *tweets*. More specifically, a tweet is a text based post of 140 characters or less, and provides information about a user's current thoughts, intentions and activities. The person making the post is called a tweeter, and other users (or tweeters) who have subscribed to the posts of a particular user are known as followers of the user. A tweet is a classic example of micro-blogging, being much smaller in size than a typical blog, but catering to topics that range from very simple to complex, thematic ones.

Twitter users have the option of restricting delivery of their tweets to those in their circle of friends, or allow everybody to access them (latter being the default case). According to a survey, 90% of tweeters keep their profiles and tweets public, thereby exposing a vast amount of Twitter data over the Internet.

Recently, Twitter developed a search engine that works on top of public user tweets. This search engine allows one to issue search queries over the last one month's tweets. At the same time, Twitter exposes a search interface in two formats – one in ATOM and the other in JSON, that lets a third party desktop or Web based application to integrate twitter's search functionality. We use the ATOM-based API to collect tweets over a period of four weeks starting from March 2009 to April 2009. In specific, our data collection module was developed using an ATOM feed reader that subscribes to tweets during this period. In order to make

Table 2: Top 20 words (based on frequency of use)

word	frequency(%)	word	frequency(%)
time	3.63	read	1.13
today	2.94	watching	1.13
come	2.03	tomorrow	1.08
try	1.65	happy	1.02
show	1.64	feel	0.97
home	1.49	fun	0.96
tonight	1.49	watch	0.91
week	1.38	live	0.87
check	1.27	video	0.86
getting	1.27	music	0.81
morning	1.22		

these tweets suitable for processing, every tweet was dumped as a triple: <user name, tweet, time of publishing the tweet>. Finally, to capture the representative set of users, we selected the most active users (i.e. those who tweeted at least once in a week) spanning across ten cities in the world. These ten cities amount for the top Twitter activity in terms of number of users [8]. Further, as per estimates [7], these users contribute towards 80% of the population of active Twitter users worldwide. Table 1 gives a synopsis of the Twitter user population in cities included in our analysis. We consider a total of approximately 21 million tweets for subsequent analysis.

3. MINING REAL–TIME USER INTERESTS

For the purpose of our analysis, we are interested in tweets of a user that capture his/her real-time interests in activities. Such tweets can be usually characterized to have one or more of the following properties:- (i) *ephemeral:* i.e. the interest in an activity changes over time as opposed to being static (e.g. hobbies), (ii)*descriptive:* the interest can be described using one or more indicative keywords or terms, and (iii) *localized:* i.e. the interest (or activity) is usually associated with (contextual) location information.

Our broad objective is to analyze tweets and identify user interests (in activities) characterized along these dimensions. In order to do so, we identify tweets expressing interests with the help of certain *content-indicative* and *usage-indicative* keywords that are present as part of the description. Content-indicative keywords (also known as category words) express the broad class (category) of user interests, e.g. movie, sports, etc. Usage-indicative keywords, on the other hand, characterize the activity associated with a particular interest. In particular, these can be either *temporal* keywords (e.g. evening) or *action* keywords (e.g. watch) that specify the time and activity associated with the category word, respectively. Note that, the location dimension of an interest can often be obtained (relatively) easily and precisely (e.g. from assisted GPS type devices, E911 type systems or the user profile itself). However, when the location is part of an interest description (e.g. downtown), one has to consider extracting this information from the tweet. For this paper, we focus on interest categories and their temporal/action–indicative keywords for the analysis of real–time interests.

To begin, we first explore what kind of keywords tweeters use most frequently and whether they contain content-indicative and/or usage-indicative keywords that could help identify user interests. Table 2 shows the results of top 20 words in our data set. We observe that there are quite a few keywords that indicate real-time user interests. These are typically nouns (describing either the type of interests e.g. game, or the time associated with the interest e.g., tonight) or verbs (e.g., watch). We exclude pronouns, prepositions, helping verbs, question words, as well as a set of non-indicative words (e.g. just, so, have) from the dataset. Since we are not interested in grammatical context or latent semantic content, we further stem the words using Porter-stemming algorithm [18] (this removes any inflections from the dataset). Next, we analyse the filtered data set for occurrences of content and usage–indicative keywords.

3.1 Content-indicative Keywords

In order to efficiently mine the tweets, we first come up with an initial list of category keywords. We consult Wordnet[2] and IMDB[3], as well as other city-specific event portals to form the list. We further enriched our seed list of keywords by manually inspecting thousands of tweets and including "interest–indicative words" observed therein. Based on empirical observations of the tweets, we finally identify five 5 *seed* categories, *viz. movie, music, food, sports, dance*, from the list of category keywords. The category keywords selected in this seed set intuitively refer to the broad class of interests (and activities) that could potentially appeal to tweeters.

Table 3 gives a snapshot of the seed categories and a few keywords corresponding to each of the seed categories.

Table 3: Category and indicative keywords describing real-time interests

Category	**Words**
Movie	movie,cinema,film,flick,sci-fi,theatre,...
Dance	dance,salsa,jazz,ballet,lyrical,disco,...
Music	music,rock,fusion,pop,rap,song,...
Food	food,restaurant,dinner,lunch,grub,...
Sports	sports,game,rugby,soccer,football,ski,...
Action	**Words**
Verbs	Watch, watching, see, view, catch,grab,get,go,play, cook,sing,drive,read, write,party,....
Temporal	**Words**
Time	tonight,today,tomorrow,evening,weekend,...

Our list contains approximately 750 category words in total including the commonly used synonyms of these terms. We have not considered hypernyms and hyponyms in this set. The list is subsequently used by Lucene[4] for gathering frequency-based statistics of stemmed category keyword usages from the universal data set of tweets. Table 4 shows the top words used in the overall dataset and percentage of total tweets containing these words.

We observed fairly significant *individual* representations of certain category keywords, such as 'game' or 'music' belonging to the seed categories discussed in Table 3. Overall, considering all category keywords, we observed that about 20% of total tweets (across the ten cities) contain atleast one keyword (as shown in Table 5). In other words, about a quarter of all tweets carry conextual information to suggest real-time user interests. Given the wide range of topics that people tweet about, these percentages are

[2]http://wordnet.princeton.edu/
[3]http://www.imdb.com
[4]http://lucene.apache.org/java/docs/

significantly encouraging for us to try and gain deeper insights into the nature of activity-oriented interests.

Table 4: Categories keyword frequencies

Cities	Content-indicative keywords (% of total tweets)
Atlanta	game (0.92), music (0.92), song (0.82), eat (0.79), food (0.54), lunch (0.49), rock (0.47), dinner (0.46)
Austin	music (0.93), game (0.87), eat (0.87), rock (0.73), food (0.71), lunch (0.70), song (0.61), dinner (0.57)
Boston	game (1.33), music (0.73), eat (0.72), song (0.65), food (0.63), dinner (0.55), rock (0.52)
Chicago	game (1.32), eat (0.80), music (0.80), song (0.75), food (0.61), rock (0.52), dinner (0.50), lunch (0.47)
London	radio (1.06), music (0.91), game (0.65), film (0.64), eat (0.63), song (0.61), lunch (0.54)
New York	game (0.96), music (0.96), song (0.84), eat (0.78), food (0.56), rock (0.51), dinner (0.41), lunch (0.40)
Seattle	game (1.13), eat (0.84), food (0.73), music (0.67), dinner (0.64), song (0.64), lunch (0.58), rock (0.56)
S. Fransisco	game (0.93), eat (0.75), music (0.73), food (0.71), song (0.65), rock (0.53), dinner (0.53)
L. Angeles	game (0.92), music (0.91), eat (0.88), song (0.83), food (0.77), rock (0.61), lunch (0.50), dinner (0.48)
Toronto	game (0.92), music (0.78), song (0.71), eat (0.71), food (0.56), dinner (0.51), rock (0.49), lunch (0.44)

Table 5: Percentage of total tweets containing at least 1 category keyword

Atlanta	Austin	Boston	Chicago	London
17.98	19.45	19.24	18.61	20.71
Seattle	N. York	S. Frans.	Seattle	Toronto
20.27	18.24	19.48	20.27	18.45

3.2 Usage-indicative keywords

Content-indicative keywords alone can provide little information as regards to inferring activity-oriented user interests. However, other words present in a microblog qualifying the content-indicative keywords can provide valuable insights into user interests. These are, as defined earlier, usage-indicative keywords. For example, if the word "movie" occurs along with "go" and "tomorrow", it gives valuable insight that a user is most likely interested in going to a movie tomorrow. We explore two kinds of such category keyword neighboring terms: (i) *action keywords* – terms indicating an action (e.g., go,see,look, etc.) and (2) *temporal keywords* terms indicating the temporal aspect of category or action keywords (e.g., today, tomorrow, etc.).

In order to gain insight on the type of such action and temporal words occurring in our data set, we use term frequency–based measure to estimate the occurrences of both temporal and action words in the data set, with seed categories described earlier. The frequency of usage-indicative words in the twitter data for the ten cities of consideration is shown in Table 6. We observed fairly high percentage of tweets containing action and temporal keywords such as *tonight, weekend* and *go, watch* that suggests the nature and temporal aspect of user intended activity. As we elaborate later, these words, when associated with category keywords, provide rich contextual information about user interests for an intended activity.

3.3 Context-based discovery of keywords

After our initial analysis of interest-related keywords in Twitter, we consider non-stemmed words to enrich our knowledge base of keywords (both content-indicative and usage-indicative). Stemmed data incurs a loss of information of tense, word sense as used in the embedded text. On the other hand, simple keyword-matching based retrieval would not work with non-stemmed data as its almost impossible to understand the different ways in which these words have been used in a massive repository of data comprised of millions of tweets. For example, even if we find all synonyms for a word "movie" from a dictionary, it is not possible to assert that "Watchmen" (a movie name) is also used in the same sense – unless we know that such a movie exists. Although we can enlist all the movies released till date, it is not an efficient and scalable solution.

We discover similar words (used in similar sense) by finding matches that are contextually similar to the seed dictionary words depicted in Table 3. Contextual similarity is measured by first building a context vector for each word indicating words occurring within a window of length k around the word. Next, Term Frequency-Inverse Document Frequency weights are created for the context vector, and *cosine similarity* is used to find words that have similarity above a certain threshold with the seed dictionary words. We had to manually inspect the result to include the new words in the keyword set. This is because the technique sometimes pulls up irrelevant words, due to the inherent noise in micro–blog data and the informal use of languages, therein.

Table 7: New category words discovered using context-based analysis

Category	Words Discovered
Movie	dvd,imax,remake,cartoon,sequel
Sports	ManU,league,uefa,footy,triathlon,tourney
Music	linkin,britney,artist,mtv,radio
Food	sushi,kfc,grenache,toastie,kebab
Dance	recital,rehearsal,flamenco,pointe,rumba

This revealed a host of new content-indicative words from twitter data for different categories. Table 7 shows a small snapshot of the keywords discovered corresponding to our seed categories. We observe that the new words discovered are often instances of a certain activity description. For example, people interested in music are more likely to talk in terms of artists, genres, bands, albums, etc. The technique could not disambiguate word senses (e.g., noun vs verb). So, "travel" for example was mostly occurring as a verb as opposed to an activity category. As a result of this limitation, the technique could not discover additional action verbs. For temporal dimension, we observed that our intuitive knowledge of seed words (e.g. "today","tonight", etc.) were fairly complete. Misspellings were also fairly accurately captured.

3.4 POS–based discovery of Action verbs

To disambiguate noun vs verb word senses, we use a Part-of-speech (POS) analyser [5] to extract action verbs. It identifies the role of a word in a sentence, including adjective, adverb, conjunction, interjection, noun (common and proper), verb and other POS phrases. The technique

[5]internal Natural Language Processing tool

Table 6: Percentages of Temporal and Action words in Data Set

Cities	Temporal keywords (%s)	Action Keywords (%s)
Atlanta	time (3.51), night (1.70), tonight (1.50), week (1.32), tomorrow (1.02), weekend (0.82)	go (5.58), do (5.31), work (3.64), love (3.44) need (3.02), see (2.59), look (2.49), watch (2.14)
Austin	time (3.77), night (1.94), tonight (1.82) week (1.45), tomorrow (1.22), weekend (0.89)	go (5.54), do (4.54), work (3.78), love (3.37) see (2.73), look (2.61), watch (2.23)
Boston	time (3.61), night (1.84), tonight (1.62) week (1.51), tomorrow (1.21), weekend (0.90)	go (5.29), do (4.71), work (3.58), love (3.35) look (2.57), see (2.51), watch (2.16)
Chicago	time (3.59), night (1.75), tonight (1.59) week (1.40), tomorrow (1.11), weekend (0.81)	go (5.70), do (4.87), love (3.58), work (3.50) see (2.56), look (2.52), watch (2.26)
London	time (3.54), night (1.67), week (1.58), tonight (1.13), tomorrow (1.12), weekend (0.85)	go (5.32), do (4.77), love (3.47), work (3.44) look (3.18), see (2.53), watch (2.21)
Los Angeles	time (3.57), night (1.82), tonight (1.54) week (1.22), tomorrow (1.06), weekend (0.79)	go (5.46), do (4.60), love (3.78), work (3.13) see (2.54), look (2.37), watch (2.33)
New York	time (3.35), night (1.68), tonight (1.37) week (1.17), tomorrow (0.93), weekend (0.78)	go (5.34), do (4.68), love (3.63), work (3.14) see (2.52), look (2.43), watch (2.13)
San Fransisco	time 0.0357, night (1.59), tonight (1.34) week (1.27), tomorrow (0.94), weekend (0.73)	go (4.68), do (4.17), love (3.15), work (3.10) look (2.43), see (2.35), watch (1.90)
Seattle	time (3.97), night (1.80), tonight (1.57) week (1.43), tomorrow (1.18), weekend (0.88)	go (5.46), do (4.74), work (3.87), love (3.48) look (2.77), see (2.64), watch (2.30)
Toronto	time (3.61), night (1.69), tonight (1.50) week (1.42), tomorrow (1.08), weekend (0.86)	go (5.05), do (4.67), love (3.49), work (3.35) look (2.76), see (2.50), watch (2.22)

locates and ranks the different grammatical roles of words that appear along with a set of "watch words" in the tweets. Using this, we could identify the relevant action verbs that show a high correlation with the identified category words (used as "watch words"). Correlation between a category word, *cw* and an action word *aw* is given by the following expression:

$$Correlation(cw, aw) = \frac{\#(A \cap B)/\#D}{(\#A/\#D)(\#B/\#D)}$$

where D represents the total number of tweets, A = {tweets containing the keyword "*cw*"}, B = {tweets containg the keyword "*aw*"}, and the # symbol represents the frequency of tweets in a set. Some examples of new action verbs discovered (*e.g., enjoy, shoot, make, taste, release, replay*) are shown in Table 8, ranked based on frequency of occurrence and correlation with the corresponding category keywords . These action verbs are added to the existing set of usage-indicative keywords and used in the subsequent section for exploring meaningful associations between content- and usage- indicative keywords.

Table 8: New action verbs discovered for category words

Category	action word	Frequency	Correlation
movie	shoot	49	3.1
movie	recommend	73	2.4
movie	release	53	2.1
movie	enjoy	189	1.7
food	taste	68	3.0
food	prepare	97	2.1
food	make	898	1.4
music	replay	39	1.3
sport	score	36	5.9

4. DISCOVERING ASSOCIATIONS IN USER INTERESTS

In this section, we explore the different latent semantic associations that exist between content-indicative category words and usage-indicative action/temporal words (i.e. those implied by the context in which words are used together). In particular, we are interested in revealing the different ways in which these words are used in tweets, discover associations between words, find out how words across these different dimensions of a real-time interest are related.

4.1 N-Gram Analysis

We are first interested in discovering usage statistics of co-occurring content-indicative words and usage-indicative words discovered previously, and part of our keyword set. A binary co-occurrence measure tells us how many tweets contain two given words. In the context of real-time interests, and from an intuitive understanding of the language, if an user is interested in an intention, he/she should use indicative action and/or temporal words to express his/her interests. e.g. I want to *watch* a movie *tonight*. Thus, usage of an action or temporal word along with a co-occurring category indicates a higher information gain for any intention-tagging heuristic.

We employ bigram-based analysis of category words (co-occurring with action verbs or temporal words)in our keyword set and report the type of observations. Co-occurring words in general for bigrams can be *at a variable distance* (characters) from each other in a document. However, Since a tweet is 140 characters in length, we employ a tolerance limit of 5 words and discover gappy bigrams where one word in the bigram is a category word, and the other is an action verb or temporal word.

Table 9 gives a snapshot of the percentage of content-bearing tweets having bigrams with action or temporal words for the ten candidate cities.

We observe that approximately 20% of content-bearing tweets contain action verbs, and hence possibly talk about real-time interests. Moreover, it is apparent that tweeters possibly use content-bearing words in several other senses as well. This is not unexpected as content-bearing words like movie, game, etc can indeed be used to convey a plethora of concepts (e.g. a movie review, a game report etc.). Some manual observation of these tweets (in London) revealed, for example, a lot of the sports–based postings about Manchester United and Chelsea games in London. It

Table 9: Content-bearing tweets having action or temporal bigrams, as a percentage of tweets containing at least one content-bearing word

Cities	(Action)	(Temporal)
Atlanta	22.4%	9.4%
Austin	21.7%	8.4%
Boston	20.5%	8.2%
Chicago	22.1%	8.0%
London	20.5%	7.4%
Los Angeles	23.0%	8.0%
New York	21.6%	7.2%
San Fransisco	19.3%	7.2%
Seattle	21.5%	8.2%
Toronto	20.9%	8.0%

Table 10: Category-wise breakdown of content-bearing tweets having action verb (bigrams), as a percentage of tweets containing at least one content-bearing word

Cities	Movies	Sports	Dance	Music	Food
Atlanta	0.8%	3.4%	0.3%	10.1%	7.8%
Austin	1.1%	3.4%	0.2%	8.3%	8.7%
Boston	0.9%	4.0%	0.3%	7.9%	7.3%
Chicago	0.9%	4.1%	0.3%	8.8%	8.0%
London	1.8%	3.3%	0.2%	8.2%	7.0%
Los Angeles	2.0%	3.4%	0.4%	9.4%	7.8%
New York	1.0%	3.3%	0.3%	9.7%	7.3%
San Fransisco	0.9%	3.1%	0.3%	7.5%	7.5%
Seattle	1.0%	4.0%	0.2%	7.9%	8.4%
Toronto	1.1%	3.4%	0.2%	8.8%	7.4%

Table 11: Tweets having <action-temporal> bigrams

Cities	(Action & Temporal)
Atlanta	5.39%
Austin	6.18%
Boston	6.02%
Chicago	5.76%
London	5.75%
Los Angeles	5.39%
New York	4.93%
San Fransisco	4.91%
Seattle	6.13%
Toronto	5.86%

is interesting to note the similarity of statistics across the candidate cities - thereby implying a uniform trend in the manner in which people tend to express interests in terms of a category (content-bearing word) and a action/temporal qualifier. Bigrams containing temporal words are lower compared to those containing action verbs. This implies that temporal terms possibly act as further qualifiers (after action verbs) to specify the context of content-bearing words - a helpful insight for further analysis of these tweets.

Table 10 breaks down the associations of action verbs across representative category words. We see maximum association of an action verb with the music category, followed by food and sports. We conjecture that the variance in these associations can be attributed to the way in which people express interests in the respective categories. For example, tweeters usually express an interests in a movie by specifying the movie name (rather than a category descriptor), thereby leading to much lower associations.

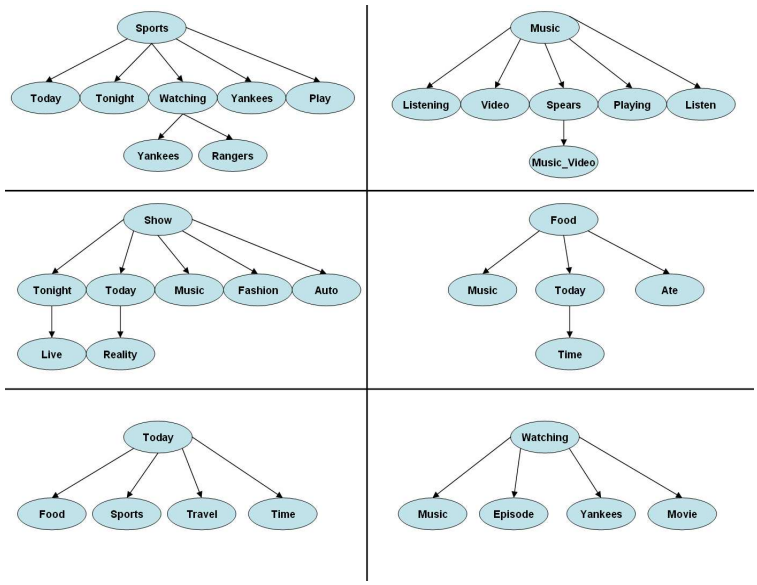

Figure 1: Tree-view of word associations discovered in relevant clusters

Table 11 reveals the percentage of tweets in the dataset that contain purely action-temporal pairs (with or without content-indicative keywords). There exists an approximate 6% of such tweets in the worldwide data . Once again, this points to the fact that people have the tendency to tweet about activities (actions) that are planned at different times of the day (temporal) - for example, *party tonight*, and hence these tweets can be exploited to uncover their short-term (near short-term) interests.

Finally, Table 12 illustrates a detailed view of the different category-action-temporal associations in each city. For sake of brevity, we show the most popular (in terms of frequency) category-action, category-temporal and action-temporal bigram associations across different cities. Quite a few of the bigrams of category words with action and temporal words are indicative of a real-time interests latent in the tweets (e.g. [“game”, “watch”], [“dinner”,“tonight”],[“go”,“eat”]). As mentioned earlier, we also observed bigram associations of action verbs with temporal words (pointing to a real-time interests) in this list. Some examples include (“go”,“tonight”), (“see”,“tonight”), (“go”,“weekend”), and (“work”,“tomorrow”).

To understand whether associations discovered between content-bearing words and action/temporal words actually signify real-time interests, we manually went through about 1000 tweets having co-occurring pairs of (“movie”,“watch”) and (“show”,“tonight”) and were encouraged to find about 25% and 33% of tweets actually meaning an intention of *watching a movie* or *going for a show*, respectively.

4.2 Contextual Analysis using k-means clustering

Our next objective is to discover any new groups of tweets that are semantically associated with real-time interests, but got omitted while forming initial categories. Secondly, we would like to perform a contextual analysis of the content-indicative and usage-indicative keywords, and understand the roles these words play in the dataset. Clustering is

Table 12: City-wise view of top cross-dimension bigrams observed in Twitter

Cities	Content word-Action verb bigrams	Content word-Temporal word bigram	Action verb-Temporal word bigrams
Atlanta	(song,love),(game,watch), (game,go),(eat,go),(music,listen)	(game,tonight),(dinner,tonight), (lunch,time),(game,time),(eat,time)	(go,time),(go,tonight), (work,time),(go,tomorrow),(see,tonight)
Austin	(eat,go),(lunch,go), (lunch,eat),(game,go),(blog,read)	(dinner,tonight),(lunch,time), (game, tonight),(game,time),(live,time)	(go,tonight),(go,time), (work,time),(go,tomorrow),(see,tonight)
Boston	(game,go),(game,watch) (blog,read),(song,love),(foot,eat)	(game,tonight),(dinner,tonight) (game,night),(game, tomorrow),(lunch,time)	(go,tonight-0.9%),(go,time-0.86%) (work,time-0.67%),(watch,time-0.384%)
Chicago	(game,go-0.57%),(game,watch) (eat,go),(food,eat)	(game,tonight),(dinner,tonight) (game,night),(lunch,time),(eat,time)	(go,tonight),(go,tommorow) (go,week),(work,week),(see,tonight)
London	(film,watch),(song,love) (radio,love),(radio,listen),(lunch,go)	(lunch,time),(dinner,tonight) (blog,time),(radio,time),(eat,time)	(go,time),(go,tonight) (go,tomorrow),(work,time),(do,time)
Los Angeles	(song,love),(game,watch) (game,go),(eat,go),(live,go)	(lunch,time),(dinner,tonight) (game,tonight),(live,time),(game,time)	(go,time),(go,tonight) (do,time),(see,tonight),(sleep,night)
New York	(song,love),(game,watch) (game,go),(music,listen),(eat,go)	(music,time),(game,tonight) (dinner,tonight),(game,time),(live,tonight)	(go,time),(go,tonight), (see,tonight),(do,tonight),(go,week)
San Francisco	(game,go),(song,love) (game,watch),(music,listen)	(dinner,tonight),(game,tonight) (lunch,time),(dinner,night)	(go,tonight),(do,time), (watch,time),(go,night)
Seattle	(game,watch),(song,love) (song,love),(blog,read)	(dinner,tonight),(game,tonight) (lunch,time),(dinner,time)	(work,time),(go,tomorrow), (watch,time),(go,hour),
Toronto	(song,love),(game,watch) (game,go),(music,listen)	(game,tonight),(dinner,tonight) (live,time),(game,night)	(go,tonight),(do,time), (see,tonight),(work,week)

a better accepted technique to group similar documents (tweets) than occurrence based filters. We used k-means clustering [6] to group relevant tweets together, while assisting the cluster formation using synonyms of content-indicative category words. We further analyze these clusters to discover latent associations of cluster tags with other words in the cluster. During clustering, we supply the tool with a set of "stop words" that include irrelevant words(e.g. prepositions, conjunctions, select adjectives and irrelevant proper nouns like http, facebook) that we do not want to observe as part of our analysis. The k-means algorithm considers all synonyms as semantically similar. This allows the tool to gather all tweets containing semantically similar words in one cluster – which, in turn, helps in revealing associations of words to the semantic interpretation of a tweet. We initially run k-means clustering with a dictionary size of 12,000 words and k = 200 as the number of clusters. The clusters were tagged with the highest occurring words (e.g. "dinner","game","movie" etc), helping us to identify clusters where we potentially uncover real-time user interests. We iterate a few times and modify the k value, dictionary size to avoid irrelevant clusters - i.e., wherein the tweets do not seem to carry a sense of real-time user interests. We next present our main findings.

4.2.1 Cluster Analysis

We discovered several new and interesting clusters across the cities (e.g. Travel, Shopping, Party etc), whose tags indicate real-time interests being captured. This enhances our characterization of content-bearing categories observed in Twitter (Table 13). We also observe clusters formed around action verbs and temporal words. For example, "watching", "weekend" came up in almost all cities as clusters. The relatively high distinctness observed in these clusters indicate unambiguous usage of these words.

We also observed a few geo-specific clusters e.g. London had "conference" as a cluster while "party" was observed in New York. Further, we observed some category-indicative clusters having a low cohesion and high distinctness (e.g. "Movie","Travel") - indicating that people use different word sets to talk about these categories, however their usage is mostly limited to these categories. For clusters that have low distinctness (e.g. "Food" in New York), we observed that they overlap with other (content-bearing) categories. Our sub-cluster analysis sheds some light on the same.

[6]We employed an internal data mining workbench containing inbuilt features for cluster analysis.

Table 13: Results of K-means clustering in two cities

Cluster	**size(%)**	**cohesion(%)**	**distinctness(%)**
Food (New York)	3.76	56.51	29.54
Food (London)	4.6	54.57	83.35
Music (New York)	3.63	53.30	30.69
Music (London)	3.51	58.54	31.21
Sports (New York)	1.66	46.33	81.40
Sports (London)	2.24	50.18	30.36
Movie (New York)	1.0	42.62	92.38
Movie (London)	1.07	48.18	86.84
New Content indicative Clusters			
Travel (New York)	1.74	47.91	78.28
Conference (London)	0.77	48.18	82.85
Party (New York)	0.93	46.81	93.35
Clusters around Action, Temporal words			
watch	0.50	47.37	91.04
read	0.50	50.15	87.64
tonight	1.89	50.52	88.75
weekend	0.63	51.82	95.04

4.2.2 Sub-Cluster Analysis

We analyzed the content of a few clusters having content-indicative tags, temporal words and action words to uncover associations *within* the clusters. We ran k-means on these clusters and gathered predominant sub-clusters. A few select sub-clusters are illustrated in Fig. 1. The results reveal (expected) implicit associations of content-bearing words with action verbs and temporal words. More importantly, it reveals which actions (or temporal words) are associated with which content-bearing words. For example, words in "music" category shows associations with

"playing" and "listen", while words in "Sports" category show associations with temporal words ("today","tonight" etc), apart from action verbs like "watching". This provides useful insight on the nature of these tweets and what can be possibly inferred from them.

For New York (and a few other cities), we observe interesting overlaps between "food", "music" and "party" categories. We find "food" present in "party", "music" present in "food" cluster. This explains the low distinctness of "food" cluster in New York, mentioned earlier. It also possibly means that people tweet about food, music, party in a non-exclusive way. Finally, for some clusters, we observe celebrity names, popular event names, names of clubs, etc. with no clear sign of real-time interests. It would require further investigation to understand some of these ambiguities and extract tweets containing real-time interests.

4.3 Temporal Analysis

Real-time interests have a significant temporal component, which if captured can lead to insights on word associations with temporal aspect of interests. Temporality can be very fine-granular. However, for real-time interests, its reasonable to expect a week-wise granularity (e.g. weekday, weekend, day of the week) and intra-day (morning, afternoon, evening, night) granularity. In the previous sections, we have presented analysis of such temporal words occurring in the data set. Here we present usage content-indicative category terms on a temporal scale, observed over a month, from March 27, 2009 to April 27, 2009.

Figure 2 shows the temporal variations observed for four content-bearing category words. We observe interesting temporal variations. We observe a spike in the sports and movie related tweets over the weekends in London, except the Good Friday weekend. Somewhat unexplained is the spike on the sports category on the Wednesday following Good Friday. Music on the other hand does not apparently show any such quasi-regular periodicity. On an average, Los Angeles users tweets more about movies than London users. Interestingly, we observe a strong spike on the food category around April 10, i.e., on Good Friday! These associations reveal additional context associated with these words in free text microblogs - which can be further exploited by diverse applications to form inference models for real-time activities.

5. PATTERN DISCOVERY IN INTEREST CLUSTERS

Based on our analysis of the data so far, we next do a microscopic analysis of select content-indicative clusters where there is a reasonable probability of extracting real-time interests. To achieve this, we built a set of benchmarks of 5000 tweets comprising of a mix of tweets from *party*,*food*,*sports* and *movie* clusters, manually tagging those that indicate a real-time interest (i.e. positive tweets). The remaining tweets in the cluster are negative tweets (i.e. tweets not indicating a real-time interest).

Next-generation social networking applications based on user interests and their activities need to exploit patterns that can efficiently and scalably extract real-time interests from such high-volume unstructured feeds. To this end, it is imperative for any classification algorithm to understand the patterns present in the tweets.

5.1 Patterns in Real-time Interest Tweets

Patterns can be of several types:

1. word occurrence-based (e.g."gym" occurs with "go" in positive tweets)
2. grammar-based (e.g. party is preceded by a verb of the form "going for" in positive tweets)
3. precedence-based (e.g. "tonight" succeeds "movie")

While benchmarking, we manually went through some relevant clusters and observed some predominant usage *patterns*. We report the different such patterns observed in the microblog data corpus.

Sports Category: Common patterns in tweets about an **intention to play a sport or go and watch a game** were:

1. going for <sport name>
2. <gonna>/<wanna> go <sport name>ing (e.g. gonna go swimming)
3. <anyone>/<who is in> for <sport name>
4. off to <sport name>ing
5. <anyone>/<who is in> for <sport name>
6. off to play <sport name>
7. heading off to <sport location> (e.g. heading off to swimming pool/stadium)

Food Category: Common patterns in tweets expressing a **real-time intention of having a food, going to a restaurant etc** were:

1. <heading out> to <dinner>/<lunch> at <place> with <someone>
2. going for <dinner>/<lunch> to <place> <tonight>/<today>
3. Dinner at <place> <tonight>
4. Dinner in some <restaurant>

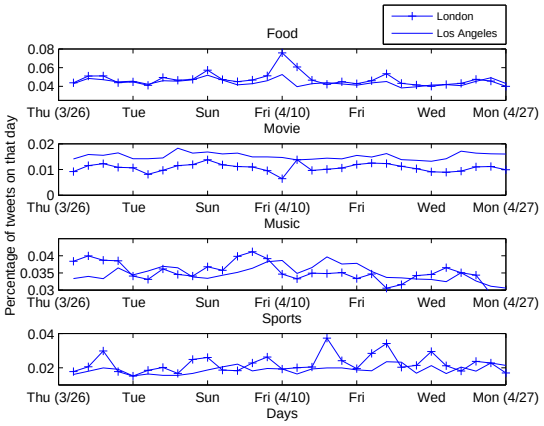

Figure 2: Temporal variation of content-indicative terms

5. looking forward to <chinese>/<indian>food at <some restaurant> <tonight><today>

6. Dinner out tonight

Party Category: Common patterns in tweets depicting user's **intention to get involved in a party** were:

1. party time (simple declaration)

2. multiple utterances of the word "party"

3. with time declarations like "party <today>/<tonight>"

4. with location specifics (e.g. "having a party at <place>")

5. others like "preparing for party <today>/<tonight>".

Movie Category: Common patterns in tweet with users expressing an intention *to watch a movie in near future* are:

1. <going for> <movie/movie name> <evening/ tonight/ tomorrow/ this weekend>

2. <movie/movie name> <tonight>

3. <time for> <movie>

4. <off to catch> <movie name> at <theatre (frequently IMAX>

5. <taking kids to> <movie/movie name> <evening/ tonight/ tomorrow/ this weekend>

Quite a few of these patterns perhaps can be represented using a combination of word occurrence-based rules or grammar-based rules. Note that, occurrence-based rules capture a reasonable number of false positives as well. Grammar-driven rules, on the other hand, might pose a scalability issue for high-volume streaming tweets. While the choice of the appropriate rule engine is beyond the scope of this paper, it is encouraging to find commonalities across the positive tweets - a helpful insight for developing online classification tools for microblogs.

We also observed a number of *conversational* tweets centered around content-bearing words in our list. For e.g. there were several tweets commenting about an ongoing event where they give updates of a game and express their feelings. These kinds of tweets can give valuable insight on the user's longer term interests. Further, it also opens up the prospect of using indirect inference to predict the user's intention (e.g. If the user is tweeting about a game, most probably he is watching it). In the party category, we observed that most of the *real-time interest bearing* tweets were posted not more than one day prior to the actual party (lot of use of temporal qualifiers like "tomorrow"). These provide important insights for a statistical classifier to build models of classifying real-time interests.

5.2 Differentiating Intentions from Tweets - Word Affinity measure

We compute the associations of frequently used words in the positive tweets and negative tweets using *affinity* as the metric. The *affinity* of a word "w" to a Set of Tweets "T", is defined as the probability of "w" to occur in "T". Table 14 provides a comparative affinity measure of select words between the two sets.

Table 14: Affinity Measures of Select Words across Positive Tweets and Negative Tweets

Category	Word	Affinity with Positive Tweets	Affinity with Negative Tweets
Movie	seeing	0.909090909	0.090909091
Movie	watch	0.857142857	0.142857143
Movie	tonight	0.716981132	0.283018868
Movie	week	0.295454545	0.704545455
Movie	film	0.302325581	0.697674419
Food	party	0.842105263	0.157894737
Food	friends	0.818181818	0.181818182
Food	beer	0.416666667	0.583333333
Party	time	0.643835616	0.356164384
Party	join	0.833333333	0.166666667
Party	nite	0.8	0.2
Sports	play	1.0	0
Sports	time	0.860082305	0.139917695
Sports	playing	0.931034483	0.068965517
Sports	watch	0.8	0.2

We observed interestingly that usage of quite a few words show a high affinity (i.e. above 0.8) towards either the positive or negative tweets. Wherever present, such high-affinity words can act as good seeds for developing frequency-based rules for fast classification of streaming tweets. We also observe variations across categories. For example, there were fewer words showing high affinity in the Food category, to either of the positive/negative classes. On the other hand, movie and sports very good separation of high-affinity words w.r.t. these classes. Another interesting observation is that a number of bigrams (e.g."cinema, tonight") came up as having high affinity towards positive tweets. This validates our initial understanding of cross-dimension bigrams having higher chances of indicating real-time interests.

5.3 Real-time Interest Classification - Initial Evaluation

Finally, using our benchmarked tweets, we performed an initial evaluation of how some traditional text classification algorithms perform in classifying positive and negative intention tweets. We trained the classifiers with 2000 benchmarked tweets (with equally biased positive and negative tweets) from the sports cluster. The tweets were selected randomly across our observation window of a month. Thereafter, classification was performed on a randomly selected work bench containing an equally biased set of positive and negative tweets. Table 15 presents the *accuracy* measures obtained in our initial evaluation.

Table 15: Accuracy Measures of well-known Text classification algorithms for extracting Real-time interests

Algorithm	Accuracy (%)
Naive Bayes	63.83
Rule Based Classifier	67.25
Decision Tree	65.37
Centroid	64.18

Initial accuracy results indicate that efficient classification of streaming tweets by social networking applications further need to exploit several mechanisms - e.g. word-usage based heuristics, rule-based filtering to classify tweets efficiently in soft-real time. Further, based on the Twitter dataset,

we observed a high degree of variation in the concepts, intentions and topics being talked about in such microblogs. This presents a huge scope for further benchmarking and exploring improved and scalable classification techniques, suitable for next-generation social networking applications.

6. RELATED WORK

Mining of user intention from natural language based dialogues have been a classical problem, researched extensively by computational linguists. A number of works [10, 20] exist on the recognition of user intentions from dialogues between people and computers in co-operative task oriented environments. A natural progression of these works was to apply the techniques to mine user generated textual data obtained from either specialized application databases [17] or from the World Wide Web [14]. However, most of the database related works focussed on the problem of document classification [16] or finding patterns and rules from the data [17], which proved particularly helpful in deducing user intentions from informative data such as those generated in a call center [13].

More recently, these techniques have been applied to analyze data obtained from various rich, informative sources in the WWW, such as the social networking sites [19], blogs [11]. The results of such analyses, however, reveal relatively static trends in user intentions and interests. In contrast, mining microblogs proved [15] to reveal more real-time based user interests and intentions, which forms the premise of our work. Although a diverse range of properties about Twitter are reported to in [15], we take a more focussed mining approach to discover real-time user intentions of participating in activities.

7. CONCLUSION AND FUTURE WORK

Motivated by the need to capture real-time user interests for an activity-oriented social network, we investigated and evaluated microblogs (Twitter) as a potential source of such contextual information about its users. An initial pass through millions of tweets collected from ten cities across the world revealed the presence of enough indicative keywords that express meaningful user interests. This seed set was then used to generate an exhaustive list of keywords, and statistical techniques applied to discover associations between these words. Further, clustering techniques revealed a number of words (and their co-occurrences) that are indicative of user interests. The patterns discovered (in terms of words and their usage) were finally validated against a set of benchmarked tweets, thereby revealing a reasonable accuracy in terms of classifying user tweets w.r.t. the context (i.e. interests, intentions, and activities) that is embedded therein. The insights obtained in the paper provide significant justification for contextual analysis of micro-blogs. We believe that there exists ample scope for research in terms of extracting rich context from social media sites – e.g. identifying user context such as emotions, presence, location etc. With the rapid evolution of mobile technologies, social media is poised to generate huge volumes of real-time consumer data. Such data presents a unique opportunity for sophisticated analytic tools and next-generation "social" applications. Our work is a step in the direction of exploring this opportunity.

8. REFERENCES

[1] IP Multimedia Subsystem (IMS); Stage 2, Release 8, 3GPP Specification TS 23.228, 2008.

[2] Parlay X Web Service Specification, Version 3.0, `http://portal.etsi.org/docbox/TISPAN/Open/OSA/ParlayX30.html`

[3] "Telco Web 2.0 Mashup: A New Blueprint for Service Creation," *Lightreading's Services Software*, Volume 3, Number 2, May 2007.

[4] Ning `http://www.ning.com/`

[5] Plazes `http://plazes.com/`

[6] Twitter `http://twitter.com/`

[7] Twitter Users Statistics `http://wiki.answers.com/Q/How many people are on Twitter`

[8] Twitter City Grader `http://twitter.grader.com/top/cities`

[9] XMPP `http://xmpp.org/`

[10] J. Allen. Recognizing intentions from natural language utterances. *Computational Models of Discourse*, pages 107–166, 1983.

[11] G. Attardi and M. Simi. Blog Mining through Opinionated Words. In *The Fifteenth Text REtrieval Conference Proceedings (TREC)*, 2006.

[12] Citation removed for double-blind review.

[13] A. K. Chalamalla, S. Negi, L. V. Subramaniam, and G. Ramakrishnan. Identification of class specific discourse patterns. In *ACM Conference on Information and Knowledge Management*, 2008.

[14] Z. Chen, F. Lin, H. Liu, Y. Liu, W.-Y. Ma, and L. Wenyin. User Intention Modeling inWeb Applications Using Data Mining. *World Wide Web: Internet and Web Information Systems*, 5:181–191, 2002.

[15] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2008.

[16] R.-L. Liu and Y.-L. Lu. Incremental Context Mining for Adaptive Document Classification. In *SIGKDD*, pages 599–604, 2002.

[17] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. *IBM Systems Journal*, 40(4):967–984, 2001.

[18] C.J. van Rijsbergen, S.E. Robertson and M.F. Porter. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587), 1980.

[19] M. Smith, V. Barash, L. Getoor, and H. W. Lauw. Leveraging social context for searching social media. In *Proceeding of the 2008 ACM workshop on Search in social media*, pages 91–94, 2008.

[20] P. Strawson. Intention and Convention in Speech Acts. *The Philosophical Review*, 73(4):439–460, 1964.