Hunter H. Janes. #Precision: An Exploration of the Utility of User-Generated Metadata for the Creation of Precise Microblog Query-Expansion Systems. A Master's Paper for the M.S. in I.S. degree. April, 2013. 34 pages. Advisor: Jaime Arguello

Twitter research provides a unique opportunity to answer fundamental questions regarding the best methods for the large-scale retrieval of extremely sparse documents. This study examines the utility of user-generated metadata expansion candidate terms for the creation of precise microblog search engines. Several search engines were created utilizing different genres of candidate expansion terms, confidence thresholds, and document parameters to explore this issue. This study demonstrates that user-generated metadata has utility for the precise retrieval of terse queries with high levels of associated conversation, such as movie awards or current events, but performs poorly on textually rich queries with lower levels of perceived conversation.

Headings:

Search Algorithms

Web Search Engines

Querying (Computer Science)

Precision (Information Retrieval)

Microblogs

#PRECISION:

AN EXPLORATION OF THE UTILITY OF USER-GENERATED METADATA FOR THE CREATION OF PRECISE MICROBLOG QUERY-EXPANSION SYSTEMS

By

Hunter H. Janes

A Master's paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
April 2013

Approved by

_____

Jaime Arguello

Table of Contents

Introduction

Twitter is the world's most popular microblogging platform. With over 500 million new messages being created on the service each day and almost 1.6 billion queries issued to its search-engine every twenty-four hours (Dugan 2012; Twitter Search Team 2011), Twitter research provides a unique opportunity to answer fundamental questions regarding the best methods for the large-scale retrieval of extremely sparse documents. This master's paper examines the structure of Twitter microblogs to determine which genre of term presents the most potential utility. A particular focus of this paper is the performance of user-generated metadata in the retrieval of microblogs. This paper specifically focuses on the usage of these terms for query-expansion. Several search-engines were created which utilized candidate expansion terms from various genres of terms, as well as different term confidence thresholds and feedback document parameters.

Potential limitations to the generalizability of this study are discussed at length below. The evaluation metric utilized for this experiment was precision, which was selected as it was assumed to best simulate the heuristic evaluation of a user. Precision was measured at five, fifteen, and thirty retrieved messages. Unexpanded queries were used as a baseline for comparison.

Literature Review

Microblogs share many of the features of a traditional blog—informality, interactivity, high frequency of update, and a standardized template platform—however, Microblogs are distinct as their entries are much shorter and thus considerably sparer. This textual sparseness lends microblogs their "micro" aspect. This scarcity of textual representation has led to a unique array of technical approaches to microblog information retrieval. The critical intervention of this master's paper is to ascertain which genre of pseudo-relevance expansion term best aids precision at *N*. Before discussing the specific motivation for this line of research, time will be spent discussing key issues in microblog research. A particular focus will be paid to how an evidence-based understanding of term quality and genre can complement a number of microblog retrieval approaches. The specific microblog investigated by this paper is Twitter.

Twitter is a microblog platform that allows users to post an unlimited number of entries, known as "tweets", which have a maximum character limit of 140 characters. Twitter has become one of the most prolifically used microblogging platforms on the internet. In February 2012 Twitter had 500 million registered users that posted roughly 340 million tweets daily (Dugan 2012). In October of 2012 Twitter CEO Dick Costolo reevaluated this number by saying that the company estimates that the company sees roughly 500 million tweets per day (Holt, 2013). According to Twitter's own estimates,

which are two years old and likely to be quite short of current volumes, roughly 1.6 billion queries are put to Twitter search each day (Twitter Search Team, 2011). Making sense of this deluge of textually sparse documents has proven to be a fertile area for inquiry.

The microblogging environment presents a unique set of challenges and is distinct from traditional blog or indexed document retrieval. The lack of textual context and girth dramatically changes the methods of retrieval and, occasionally, the object of evaluation (Efron 2011). Scholarly approaches to microblog retrieval have largely focused on two complimentary research streams to overcome these textual limitations: the exploitation of structural information (Kalmanovich and Kurland 2009; Vieweg et al. 2010), and query expansion (Meij, Weerkamp, and De Rijke 2012). Exploiting structural information has seen wide array usage. It has been used to geo-locate messages (Vieweg et al. 2010), harvest Hashtags (Efron 2010), and analyze rebroadcast patterns to overcome the anemic amplitude of the original information signal. Other approaches have used temporal information to identify tweet-bursts to identify tweets with high expansion potential:

> For many queries, the relevant tweets are concentrated in temporal bursts in the past and that identifying these bursts and favoring tweets published during those time periods can improve retrieval. (Willis et al 2012, 1)

Some of the more successful approaches have attempted to make use of almost all available structural and temporal information when expanding a query:

> Query expansion on microblog data can be done in a dynamic fashion (taking time into account) and should include specific terms like usernames, hashtags, and links. (Massoudi et al 2011)

However, most query expansion research has focused heavily on using outside textual information such as identifying topics in messages (Vieweg et al. 2010), or using text

from semantically related messages as a form of document expansion. Methods which attempt to dramatically expand queries using outside signals are technically interesting, but have shown themselves to be difficult to implement with precision:

> Indeed, there are many queries for which state-of-the-art PF [pseudo-relevance frequency] expansion methods yield retrieval performance that is substantially inferior to that of using the original query with no expansion—the performance robustness problem. (Kalmanovich and Kurland 2009, 646)

The performance robustness problem arises from the tendency of expanded content to introduce an untenable amount of "query drift" (Kalmanovich and Kurland 2009). This drift exhibits itself as the phenomenon in which results with relevancy for a different query task than the original begin to be returned to the users. Outside and open-source document databases such as Wikipedia (Elsas et al. 2008) have proven to be favorites amongst researchers hoping to dramatically expand the content of microblog documents. Until the performance robustness problem can be convincingly addressed, using outside signals for query expansion will remain a promising though ultimately unreliable method.

Microblogs are known to be noisy (Kelly 2009). Each microblog platforms has an unique method for winnowing results, though none appear to be effective at reducing the volume of unwanted messages. Twitter is notoriously noisy and is a system with a massive volume of new information passing through it every day. Due to its 500 million daily tweets and 1.5 billion new daily queries, it could be argued that the desire to winnow information is the driver behind a majority of Twitter studies. Methods for winnowing the amount of information and maximizing quality of information are often rooted in the desire to follow conversations with a minimum of signal noise. In the

massive, rapid, and time sensitive environment of most microblogging services, following a particular conversation across time is not a solved problem.

If one limits oneself to a review of research to which has been done specifically to improve *ad hoc* retrieval methods on the Twitter platform, the most common practice has proven to be the harvesting of structural information from large volumes of tweets. This approach has been popular as Twitter's application programming interface (API) has made harvesting large numbers of messages a *de minimis* task. Several well established information retrieval researchers have attempted to exploit structural information from Twitter to overcome its textual deficiencies. These methodologies have largely been concerned with using the structure of individual tweets to gain quality information for query expansion (Kalmanovich, 2012; Vieweg et al, 2011; Meij et al, 2011). Most of this work has dealt with simple term frequency and mutual information. There has been significantly less work done specifically exploring the potential of Hashtag mining, the exploitation of user-created metadata, for conversational monitoring and query expansion. This seems to be due to the difficulty of exploiting Hashtag information for *ad hoc* retrieval.

In this work, we explore the use of hashtags for query expansion on the Twitter microblogging platform. Hashtags have seen wide acceptance on Twitter and have been adopted into the formal design of the system and are thus worthy of a focused discussion. Hashtags have their roots in a design implemented by creator Jarkko Oikarinen for Internet Relay Chat (IRC). Oikarinen's design placed a hash mark in front of a noun to denote IRC conversational channels. Hashtags were later introduced to the Twitter lexicon on August 23, 2007, when researcher Chris Messina tweeted the following

message: "how do you feel about using # (pound) for groups. As in #barcamp [msg]?"[1] Twitter has since adopted this method for creating publically findable conversations. The Twitter Hashtag is a form of user-generated metadata consisting of a Hashtag (#) followed by a string of alphanumeric characters, and has thus not diverged noticeably from Oikarinen's original design and intention. Hashtags provide a unique opportunity to utilize user-generated metadata to assist in information retrieval. Even if they prove to be completely detrimental to informational retrieval efforts, which the experimental section of this paper refutes, they are used with such frequency that they must be dealt with by any Twitter search engine.

A highly influential work for the development of the methodology of this master's paper was Miles Efron's 2010 work *Hashtag Retrieval in a Microblogging Environment*, which outlined the motivation for harvesting Hashtags in the following manner:

> "Finding useful Hashtags offers benefits that are particular to the microblog setting:
>
> - Tags to follow: A user may wish to find Hashtags that he or she can track on an ongoing basis.
> - Result display: If users are searching for tweets, people, or posted URLs, returned units of retrieval could be grouped into clusters by their associated Hashtags.
> - Query expansion: Hashtags provide leverage for query expansion during relevance feedback." (787)

This paper focuses exclusively on Efron's third criteria, but may have implications for both the first and second. Once an experimentally derived expansion threshold is discovered for Hashtags, then their pragmatic application will likely be greatly increased to the remaining criteria. There is a documented research corpus detailing the use of structural information to provide information for pseudo-relevance feedback for Twitter

[1] https://twitter.com/chrismessina/status/223115412

query expansion. As mentioned in previous pages, this approach has been popular largely due to the ease by which Twitter's application programming interface (API) can be queried to harvest large volumes of tweets. However, it should be noted that the harvesting of tweets is not completely open. There are various free and paid levels of access for registered Twitter developers. Twitter has recently revisited its harvesting policies and has shown a demonstrated tendency to limit high-fidelity access to paying researchers in recent years.

By exploiting forms of structural information such as user-created metadata, it may be possible to ameliorate some of the more detrimental aspects of query expansion methods such as pronounced query drift. Though query drift does assert itself through the use of Hashtag information, as was discovered during the analysis phase of this project, this effect can be accounted for by aggressively limiting the Hashtags used for expansion. Several previous approaches have used offline machine learning to identify clusters of semantically related Hashtags (Kalmanovich et al. 2009). Several of the more successful attempts have used various methods of bootstrapping to identify additional Hashtags to cluster (Lee, Croft, and Allan 2008). There have also been attempts to use simple term frequency (Cui et al. 2012) as a proxy of influential Hashtags and conversations. These efforts have been met with varying levels of success.

There has not been a prolonged inquiry into not only what genre of expansion terms make for the highest precision, but also what feedback document threshold provides the best mine of candidate expansion information. The addition of an investigation of what level of tag retrieval query drift begins to occur may make the approach of this master's paper novel. By tuning these parameters, it is thought that the utility of hashtags as query

expansion candidates can be explored. Furthermore, these methods can serve as basis for future projects which can use textual or structural information to combat the major problems faced by microblog retrieval. By preforming basic scientific research into this fundamental unit of investigation better results in future experiments might be obtained. The experimental design of this paper makes use of several feedback document levels and confidence thresholds to begin to address the robustness problem faced by microblog retrieval. This ultimate goal of this paper is to begin to move towards an optimal methodology for avoiding the problem of performance robustness by exploiting different genres of candidate terms.

Methodology

Experimentation was performed on the 2011 Twitter corpus provided to the the Text Retrieval Conference (TREC) Microblog track. Through the manipulation of various tunable parameters, it was hoped that the best genre of candidate expansion term for query expansion could be determined. The genres of candidate terms in this corpus were Hashtags, Non-Hashtags, and a combination of both genres. A secondary goal of this analysis was to determine which feedback document level yielded the best results. Finally, each pool of expansion candidates was subjected to a severe reduction in number of expansion terms to test for query drift. The intended outcome of this experiment was to allow a researcher to confidently know which terms to utilize for query expansion, and which feedback document level yielded optimal results. The metric of evaluation was precision at five, fifteen, and thirty. Corpus indexing, pseudo relevance feedback, and retrieval evaluations were performed using the utilities contained in the Indri toolkit, which is a tunable search engine tool developed in cooperation between the University of Massachusetts and Carnegie Mellon University as part of the Lemur Project. Textual manipulation and generic term filtering were performed using a regular expression scripts and a programmable text editing software solution.

Indri was installed in a university-run Linux-based computing cluster consisting of roughly 8000 computing cores and 706 independent servers (The University of North

Carolina at Chapel Hill, 2013). This computing framework allowed for rapid testing, evaluation, and assessment. As stated above, the corpus used was a subset of the 2011 Twitter corpus provided by the Text Retrieval Conference (TREC) Microblog track. This corpus subset consisted of roughly five million unique tweets collected during a period spanning January 23rd- February 8th 2011. The full experimental corpus consisted of 16 million unique tweets, but this researcher was unable to gain access to the full corpus due to both logistic and privacy issues. Provided with this corpus were 110 queries, covering a wide array of topics, and a full set of National Institute of Standards (NIST) evaluations. These evaluations were performed in a binary relevant/irrelevant manner by professional trained assessors.

After acquiring access to Indri, a digital copy of the corpus, and the Linux computing cluster, the five million tweets were modified using regular expression scripting. This was done to ensure that the preceding Hashtag symbol (#) was not removed by Indri during the indexing, stemming, or querying stage of the experiment. This purpose-built regular expression script transformed the symbolic Hashtag into a natural language Hashtag: i.e. the Hashtag #Syria was transformed to the term HASHSyria. A stop word list was obtained and utilized by this researcher to prevent the substitution of stop words for more meaningful terms during query expansion. The stop word list used was provided by The Journal of Machine Learning research, which is an academic journal produced and managed by the MIT Press. Once the text of the corpus was correctly transformed, and reliable stop word obtained, it was indexed using the *IndriBuildIndex* utility into an inverted index. The results of the indexing were then stored on the Linux computing cluster for experimental usage.

This index was next mined for candidate query expansion terms. This was done in multiple passes in which the top thousand terms from twenty-five, fifty, one hundred, and one thousand feedback documents were selected. The feedback documents were selected algorithmically by the Indri system, and represented the most confident documents produced by the baseline retrieval algorithm. All tunable parameters were set to their default Indri parameters: μ smoothing parameter (0) and weighting (.5). This ensured that both expansion terms and original query terms were given equal weight. These parameters were left untuned as the experimental design was intended to test *only* the best genre of expansion term and the feedback document level, and the tuning of additional expansion criteria was outside the scope of this project. The corpus was then queried using the four previously mentioned document parameters using the IndriRunQuery utility. This process resulted in the following useable expansion term counts displayed in the below figure.

Usable Terms Obtained From Expansion Harvest			
Number of Feedback Documents	**Hashtags Retrieved**	**Non-Hashtags Retrieved**	**Total Terms Retrieved**
25	506	16298	16804
50	981	31405	32386
100	1863	56371	58234
1000	3177	103772	106949

This retrieval of Hashtag terms closely followed Heap's Law as that the rate of additional terms retrieved quickly diminished as the number of feedbacks documents increased.

Thus, retrieving Hashtags from a large feedback document corpus would likely not yield significantly more candidate terms than a much smaller feedback documents number. This has implications for computational efficiency and will be discussed in the discussion section.

Once harvested, potential query expansion terms were sorted into query term pools consisting of Hashtags, non-Hashtag terms, and mixed bag of expansion terms. Of these pools, the first two (Hashtag only and non-Hashtag) were mutually exclusive pools that contained no term overlap. This was accomplished using another purpose built regular expression script.

From these pools the top twenty candidate expansion terms for each of the 110 queries were selected. These terms were then ranked in accordance to their assigned expansion weights. Each set of twenty expansion terms were then used as expansion terms for each query. As a means for adjusting for potential query drift, a separate query pool was created for each genre which made use of the most confident candidate expansion term. The most confident expansion terms was determined by the candidate term with the highest weight. This resulted in twenty-five unique queries: three separate queries for each of the four document expansion parameter pools, twelve queries based on the most confident expansion term for each document and genre pair, and baseline non-expanded query. These twenty-five queries were performed on the previously indexed corpus. The output of these queries was stored on the Linux cluster for use and evaluation.

The results of the queries were evaluated using the Indri *TRECEval* utility and the provided TREC Qrels evaluation file. The metrics used for this experiment were precision at five, fifteen, and thirty. These precision levels were set to simulate several possible uses of the system by the user. The precision of each query was the measured at all three precision levels. The precision of each query pool at each precision level was then compared and the results analyzed. After this comparison was performed, the nature of the individual 110 queries were examined. This provided the researcher with not only data about how well the genres of candidate terms preformed overall, but also how well they performed on each of the 110 queries.

Experimental Results

The results of the query expansion were evaluated for best performance at the precision levels of 5, 15, and 30 to simulate potential user interaction with a retrieval system. These were next evaluated to examine which expansion method and which feedback-document parameter gave the most precise results. This was then used to determine what expansion method and feedback document level was the best method for each precision level, as well as what method would be the best expansion method overall. The results are given in figures 1-A, 1-B, and 1-C below. The highest achieved precision is highlighted in green.

Figure 1-A

Precision @ 5		
Genre of Term	Best Precision	Feedback Document Count
Baseline	0.363	N/A
Non-Hashtag Term	0.3685	50, 100
Most Confident Non-Hashtag Term	0.3611	25,100
Hashtag Term	0.2185	25
Most Confident Hashtag Term	0.3	25
Mix Terms	0.4056	1000
Most Confident Mix Term	0.3685	50, 100

Figure 1-B

Precision @ 15		
Genre of Term	Best Precision	Feedback Document Count
Baseline	0.334	N/A
Non-Hashtag Term	0.3673	25
Most Confident Non-Hashtag Term	0.321	25
Hashtag Term	0.1796	25
Most Confident Hashtag Term	0.279	25
Mix Terms	0.3617	25
Most Confident Mix Term	0.321	25

Figure 1-C

Precision @ 15		
Genre of Term	Best Precision	Feedback Document Count
Baseline	0.288	N/A
Non-Hashtag Term	0.3235	25
Most Confident Non-Hashtag Term	0.2799	25
Hashtag Term	0.1636	25
Most Confident Hashtag Term	0.241	25
Mix Terms	0.3123	25
Most Confident Mix Term	0.2799	25

With the notable exception of the mixed terms with a feedback document count of 1,000, the best results were achieved from methods with used terms with only twenty-five feedback documents. The only methods to show any performance above baseline as the precision evaluation levels of fifteen and thirty where methods which made use of no more than twenty-five feedback documents. The most consistently stable feedback methods, as measured by performance above the baseline, where the mixed term and non-Hashtag term methods. These did not exhibit the marked tendency towards relevancy deterioration exhibited by the purely hashtag-based methods. The highest precision achieved in this evaluation was .405, which was achieved using a mixed bag of words

and a thousand feedback documents. This solution was not very robust, as it performed poorly on all precision levels. This suggests that sever query drift was introduced into the retrieval using this method. Using this corpus, and only this corpus, it can be said that the safest method of retrieval is to use twenty-five feedback documents and non-Hashtag term candidate expansion. The applicability of this methodology to a search engine on "wild" Twitter is suspect, as is discussed at length below.

It was observed during the experiment that hashtags exhibited rapid ranked relevancy decay when compared to non-hashtag terms. That is to say that a hashtag at rank $N>1$ is more likely than a non-hashtag to produce a non-relevant result in this particular microblog corpus. The standard method of expansion, in which the twenty terms most confident expansion candidates were utilized, produced poor results for pure hashtag retrieval. This is likely due to the relevant scarcity of hashtags in this corpus, and thus selecting twenty hashtags from a scare supply forces one to utilize candidate expansion terms with very low confidence values. Severely limiting the number of hashtags used in a query, through limiting the scope of hashtags used to the single most confident hashtag in the set, yielded markedly better results. Hashtags were the only genre of candidate term to benefit from reducing the expansion to the single most confident term. This is strongly suggestive that the proper usage of Hashtags would be to use a small number of extremely confident candidate terms during query expansion. It would appear that Hashtags make for fragile expansion terms, but have the ability to produce relevant results when treated with the proper care.

Moving away from conjecture and back to the results of the performed experiment, it can be said that the best results were achieved using the top twenty-five

feedback documents and using non-hashtag query expansion terms. This approach proved to be the most consistent at the most levels of precision, and did not fall victim to the robustness problem at the same rate as hashtags. This leads to the conclusion that if one were to build a system to handle ad hoc style retrieval for Twitter, then they would want to discredit hashtag terms. As stated above this is likely due to the fact that hashtags appear in very low frequencies in returns for ad hoc style queries. This is despite the fact that they appear in very high frequency on Twitter in its naturalistic setting. This brings up the problem of generalizability of findings from this corpus to actual Twitter search.

Discussion

The observed variation in the performances of the expanded queries may be partly due to the way that TREC constructed test queries, which seemed to be more aligned with traditional *ad hoc* search engine searching than with searches performed naturalistically on Twitter. This likely led to the deep variation between queries which worked well for hashtag only retrieval and queries which worked well for non-hashtag retrieval.

A specific example of this variation can be found in query #60 ("Fishing Guidebooks") which, when the feedback document parameter was set to 25, had a precision at five of .8667 when non-hashtag candidate terms were used. This same query did not break the >.5 precision level when using the hashtag candidate system. This may be due to the nature of the query. "Fishing Guidebooks" is an extremely specific query, and is thus likely not to spawn a large volume of Twitter conversation due its lack of broad and general cultural interest. Conversely, a topic with a fair amount of assumed general interest and spontaneous conversation, such as query #88 ("Kings Speech Awards"), yielded a perfect retrieval score of P@5 = 1.0 followed by P@15 = .933, and P@30= .8 when using the hashtag only retrieval system. For the same query, the non-hashtag query system's highest retrieval score for that query was P@5 = .8. This suggests that hashtags may be highly reliable for tasks which center on queries with high levels of user conversation, but perform poorly for others. The TREC corpus may favor queries

that treat Twitter, a platform built to facilitate extemporaneous conversations, as if it were a traditional document retrieval system. This currently remains an open question, until future work can more thoroughly investigate this topic.

Queries such as "Phone Hacking British Politicians" (#8) and "Dog Whisperer Cesar Millan's Techniques" (#33) are much better suited to expected user behavior for an *ad hoc* retrieval system for indexed documents such as Google or Yahoo. They appear to be much too large and much too specific for microblog usage. Query #33 is so large that it takes up almost a full 30% of the maximum size of the document, 140 characters, which is its unit of retrieval. The likelihood of any particular microblog document satisfying such a large and specific query is very low, and thus there is a strong chance that the precision of such retrieval would be quite poor no matter what system is utilized. A much more naturalistic query list and corpus might yield results that not only have higher precision, but are more generalizable to the actual system in question. It is telling that no system utilized had a precision that an average user would likely find acceptable. Hashtags may prove to be a method for retrieving these more general interest and conversational queries.

Twitter searches tend to be, much like their unit of retrieval, quite spare, broad, and conversational. The top trending searches at the time of this writing (3/21/2013 9:38 PM EST) are as follows: #PAX, #E3ChloeBoston, #WhyISmile, #ImSoUsedToHearing, A&T, #teaching2030, Martin Brodeur, Louisville, NCAA Tournament. It should be noted that no search exceeds two terms and that fully 55% of the top searches are for conversations centering on hashtags. This suggests that the naturalistic usage of Twitter is

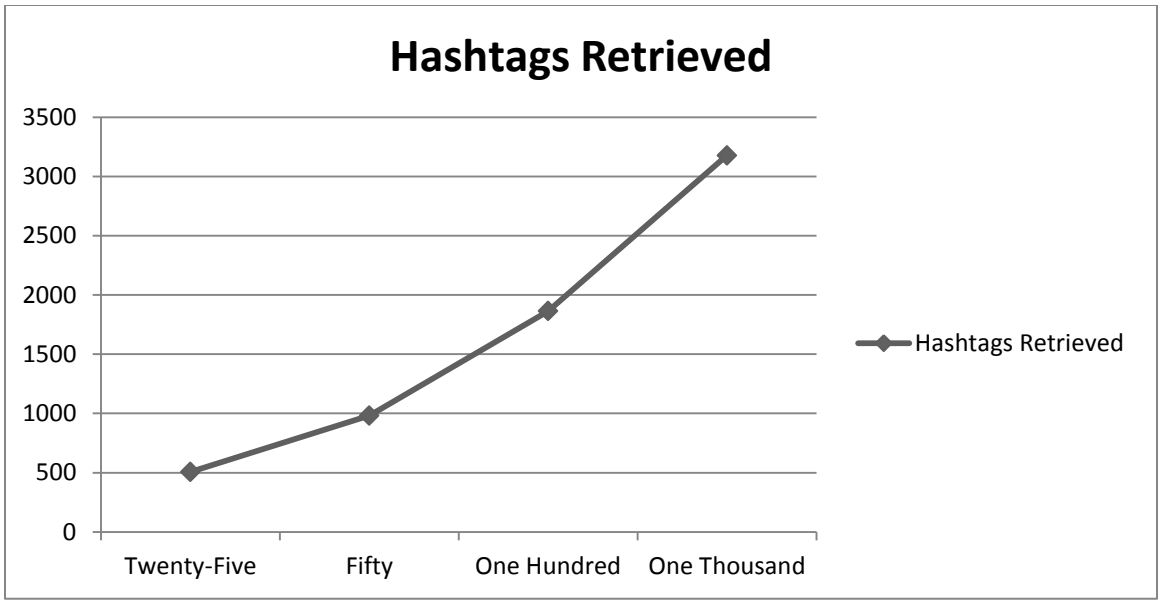
not to only search using extremely sparse queries, but to query for conversations in which to take part.

It is also interesting to observe that a large number of the hashtags in the Twitter's top searches are parlour tags. These repurposed hashtags, which function very much like Victorian parlour games, serve no notable informational purpose apart from simply participating in an ongoing quasi-game of interpretation and appropriation of the meaning of the hashtag. For example a search for #WhyISmile produces tweets such as: "#WhyISmile my zygomatic muscle is working properly", "#WhyISmile smile because it reminded me of someone that I like (◠‿◠) ", and "#WhyISmile to hide my emotions and feelings". If a search engine is too be built for Twitter, these parlor tags need to be dealt with in a rigorous and systematic manner. After all, these make for some of the most popular searches to pass through Twitter's search engine.

During the harvesting phase of the experiment Hashtags followed Heap's Law:

$$V_R(n) = Kn^{\beta}$$

Thus, the return on computing investment from harvesting Hashtags from 1000 documents instead of 100 documents is quite low. In the chart displaying this behavior below, special attention should be paid to the X-axis as the movement from 100 to 1000 documents is a tenfold increase in document pool size for only a 58% increase in number of retrieved hashtags.



It would therefore likely be a poor use of computational resources to attempt to harvest Hashtags from a large feedback document pool. This diminishing return on investment was not as steep for other genres of candidate expansion terms. This implies that a system which can successfully make use of Hashtags would have the added benefit of being able to mine less feedback documents than other systems, thus saving on computing power and time.

A thorough study of general Twitter usage will need to be undertaken to determine the generalizability of these findings. The goal of such a study would be to determine both the nature and structure of Twitter queries and usage. This would allow for the construction of a corpus for testing retrieval systems in a more naturalistic setting. Confounding the effort to create a more naturalistic retrieval corpus is the observation that Twitter is a duel-purpose system. Twitter serves both as a microblogging platform and a social network. It is distinctly different from other social networks, such as Facebook, in that users do not typically “follow” the uses that they know in their offline

lives. Twitter users often follow a wide number of accounts that span a variety of interests and genres of information. New Twitter users must engage in a complex network building process to discover users that post things of interest. Users often end up manually constructing a network by following a recursive path of discovery of interesting microbloggers from the tweets of microbloggers they are already interested in. Users are not only searching for relevant tweets, but they are also searching for relevant users. A truly comprehensive Twitter search engine would need to address not only the nature of Twitter search, but incorporate the social aspirations of its users into its design.

Returning to the specific inquiry of this paper, it is interesting to consider how a more naturalistic query corpus would respond to hashtag-only retrieval. As discussed earlier, the TREC corpus contains queries which may be unrealistic in both content and length. In addition to changing the construction of test queries and recognizing the nature of Twitter as a social network, there may need to be a reconsideration of current relevancy evaluation methods. Evaluating rather something is relevant to a conversation is quite different from evaluating rather a tweet is relevant to a specific query. As Tweet are units of conversation, and not simply micro-documents, they do not have the same binary (ir)relevancy of documents traditional utilized in *ad hoc* searches. Tweet relevancy likely operates on a much fuzzier logic. Relevancy is an elusive target, as Twitter users are constantly appropriating, redirecting, adding to, and subverting conversations. How exactly this should be taken into account when performing relevancy judgments is beyond the scope of this paper.

Ultimately, the best solution may be to bifurcate the construction and optimization of a Twitter search engine. One engine would be tuned to handle *ad hoc* style queries,

while another would be created to handle the conversational and social core of Twitter. This would ensure that both engines would be optimal for two potential usages of Twitter search, while not compromising the function of either.

Future work

The observation that hashtags perform well for general interest and conversational Twitter queries may prove to be of use to future investigations. A potential extension of this paper might be the creation of a metric to determine the relative volume of conversation surrounding specific queries, and thus ascertain rather a query is a "safe" candidate for hashtag expansion. Such an approach might function through the exploitation of structural and temporal elements of Twitter: retweet volume and frequency of query-related tweets. This expands upon recent research into tweet "bursts" to identify candidate expansion tweets (Willis et al, 2012). Queries with high levels of conversation may prove to be, if the findings of this paper prove to be generalizable, ideal candidates for hashtag-based query expansion. Filtering queries by the amount of conversation would help to ensure that queries ill-suited to such forms of expansion would be spared high-levels of unnecessary query drift.

An additional area for future work is in the area of relevancy evaluation and corpus construction. Dealing with the ambiguity of conversational relevancy and the construction of naturalistic queries are areas with deep potential for improving the field of microblog retrieval. These research streams would require a sustained look at nature of Twitter, and may thus benefit from both computational and humanistic inquiry.

As was mentioned in the discussion section, it may prove fruitful to construct a bifurcated query expansion system for Twitter. Such a system would seperate *ad hoc* queries from queries intended to discover interesting conversations or users. This would allow for the focused tuning for the performance of two unique user tasks, without sacrificing the unique expansion strengths of the retrieved corpora of either approach. Due to its reliance on the recognition of the nature of a query, and need for reliable relevancy assessment, this dual approach might be best undertaken once the previously mentioned areas for future research have been more thoroughly explored

Conclusion

It is the finding of this paper that non-hashtag terms are optimal for *ad hoc* style microblog retrieval. This recommendation is offered with some reservations as there is suggestive evidence that the task of *ad hoc* retrieval is not representative of the most popular Twitter searches. The task of conversational discovery may be a more realistic usage for a Twitter search engine, and in such cases it is the recommendation of this researcher that a mixture of properly scoped hashtags and non-hashtags be employed. Such a system would have the advantage of both user-created metadata and the numerical robustness of non-hashtag terms. Statistical modeling has revealed that Hashtags, when an appropriate confidence threshold is utilized, have the ability to offer relevant results at computationally insignificant costs due to their tendency to follow Heap's Law. Furthermore, it was revealed that Hashtags perform at much higher relevance levels when their confidence threshold is severely limited.

A complicating factor is the likely need for the development of new modes of relevance judgment. As microblog queries may be best viewed as mechanisms to discover conversations, and not simply a novel method for accessing very small documents, then perhaps other evaluation measures should be developed for relevancy judgments. The introduction of fuzzy logic or a relaxation of the standards of relevance might need to be considered by National Institute of Standards (NIST). How exactly NIST might reformulate its relevancy judgments is beyond the current scope of this paper.

It is thus the recommendation of this author that any system which attempts to utilize Twitter data for query expansion should carefully bifurcate its development efforts. This will ensure that optimization efforts which attempt to utilize candidate expansion queries will not be counter-productive.

Works Cited

Balog, Krisztian, Leif Azzopardi, and Maarten De Rijke. 2006. "Formal Models for Expert Finding in Enterprise Corpora." *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 43–50. http://dl.acm.org/citation.cfm?id=1148181.

Balog, Krisztian, Maarten De Rijke, and Wouter Weerkamp. 2008. "Bloggers as Experts: Feed Distillation Using Expert Retrieval Models." *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 753–754.

Boyd, Danah. 2009. "Twitter: 'Pointless Babble' or Peripheral Awareness + Social Grooming." http://www.zephoria.org/thoughts/ archives/2009/08/16/twitter_pointle.html.

Carter, Simon, M Tsagkias, and W Weerkamp. 2011. "Twitter Hashtags: Joint Translation and Clustering": 1–3. http://journal.webscience.org/529/.

Celik, Ilknur, Fabian Abel, and GJ Houben. 2011. "Learning Semantic Relationships Between Entities in Twitter." *Web Engineering*: 167–181. http://www.springerlink.com/index/5673031T47304276.pdf.

Craswell, Nick, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. 2001. "P@noptic Expert: Searching for Experts Not Just for Documents." *Ausweb Poster Proceedings, Queensland, Australia*. http://es.csiro.au/pubs/craswell_ausweb01.pdf.

Croft, Bruce, Donald Metzler, and Trevor Strohman. 2010. *Search Engines: Information Retrieval in Action*. Addison-Wesley.

Cui, Anqi, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. "Discover Breaking Events with Popular Hashtags in Twitter." *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12*: 1794. doi:10.1145/2396761.2398519. http://dl.acm.org/citation.cfm?doid=2396761.2398519.

Dugan, Lauren. 2012. “Twitter To Surpass 500 Million Registered Users On Wednesday.” *Media Bistro*, February 21. http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842.

D’Orazio, Dante. 2012. “Instagram Cuts Off Twitter Cards Integration, Further Souring Relationship.” *The Verge*, December 5. http://www.theverge.com/2012/12/5/3730876/instagram-cuts-off-twitter-cards-integration-further-souring-relationship.

Efron, Miles. 2010. “Hashtag Retrieval in a Microblogging Environment.” *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’10*: 787. doi:10.1145/1835449.1835616. http://portal.acm.org/citation.cfm?doid=1835449.1835616.

———. 2011. “Information Search and Retrieval in Microblogs.” *Journal of the American Society for Information Science and Technology* 62 (6) (June 16): 996–1008. doi:10.1002/asi.21512. http://doi.wiley.com/10.1002/asi.21512.

Elsas, Jonathan, Jaime Arguello, Jamie Callan, and Jaime G Carbonell. 2007. “Retrieval and Feedback Models for Blog Distillation Retrieval and Feedback Models for Blog Distillation.” *Computer Science Department* Paper 282. http://repository.cmu.edu/compsci/282/.

Elsas, Jonathan, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. 2008. “Retrieval and Feedback Models for Blog Feed Search.” *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’08*: 347. doi:10.1145/1390334.1390394. http://portal.acm.org/citation.cfm?doid=1390334.1390394.

Holt, R. (2013, March 21). Twitter in numbers. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html

Hearst, Marti A. 1999. “The Use of Categories and Clusters for Organizing Retrieval Results.” In *Natural Language Information Retrieval*, 333–373. Kluwer Academic Publishers.

Hearst, Marti A., Matthew Hurst, and Susan T. Dumais. 2008. “What Should Blog Search Look Like?” *Proceeding of the 2008 ACM Workshop on Search in Social Media - SSM ’08*: 95–98. doi:10.1145/1458583.1458599. http://portal.acm.org/citation.cfm?doid=1458583.1458599.

Jansen, B J, M Zhang, K Sobel, and A Chowdury. 2009. “Twitter Power: Tweets as Electronic Word of Mouth.” *Journal of the American Society for Information Science and Technology* 60 (11): 2169–2188.

Kalmanovich, Inna Gelfer, and Oren Kurland. 2009. “Cluster-based Query Expansion.” *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’09*: 646. doi:10.1145/1571941.1572058. http://portal.acm.org/citation.cfm?doid=1571941.1572058.

Kelly, R. 2009. *Twitter Study Reveals Interesting Results About Usage*. San Antonio, TX. http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter- Study-August-2009.pdf.

Lee, Kyung Soon, W. Bruce Croft, and James Allan. 2008. “A Cluster-based Resampling Method for Pseudo-relevance Feedback.” *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’08*: 235. doi:10.1145/1390334.1390376. http://portal.acm.org/citation.cfm?doid=1390334.1390376.

Meij, Edgar, Wouter Weerkamp, and Maarten de Rijke. 2012. “Adding Semantics to Microblog Posts.” *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM ’12*: 563. doi:10.1145/2124295.2124364. http://dl.acm.org/citation.cfm?doid=2124295.2124364.

Mishne, Gilad. 2006. “Information Access Challenges in the Blogspace.” *The International Workshop on Intelligent Information Access (IIIA 2006)*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.9533&rep=rep1&type=pdf.

Mishne, Gilad, and Maarten De Rijke. 2006. “A Study of Blog Search.” *Advances in Information Retrieval*: 289–301. http://www.springerlink.com/index/j1254g6h75747725.pdf.

Nagmoti, Rinkesh, Ankur Teredesai, and Martine De Cock. 2010. “Ranking Approaches for Microblog Search”: 153–157. https://biblio.ugent.be/publication/1108169.

K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In Proceedings of the 33rd European conference on Advances in information retrieval, ECIR `11, pages 362{367. Springer-Verlag, 2011.

Sankaranarayanan, J, H Samet, B E Teitler, M D Lieberman, and J Sperling. 2009. “Twitterstand: News in Tweets.” In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 42–51. ACM.

Sippey, Michael. 2012. “Delivering a Consistent Twitter Experience.” *Twitter Developer Blog*, June 29. https://dev.twitter.com/blog/delivering-consistent-twitter-experience.

Streams, Kimber. 2012. “Twitter Takes on Third-party Developers with Strict New Rules.” *The Verge*, August 23. http://www.theverge.com/2012/8/23/3263481/twitter-api-third-party-developers.

Twitter Search Team. (2011, May 31). *The engineering behind twitter’s new search experience*. Retrieved from http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html

Tsur, Oren, Adi Littman, and Ari Rappoport. 2012. “Scalable Multi Stage Clustering of Tagged Micro-messages.” *Proceedings of the 21st International Conference Companion on World Wide Web - WWW ’12 Companion*: 621–622. doi:10.1145/2187980.2188157. http://dl.acm.org/citation.cfm?doid=2187980.2188157.

The University of North Carolina at Chapel Hill. (2013, February 27). *Getting started on killdevil*. Retrieved from http://help.unc.edu/help/getting-started-on-killdevi

Vieweg, Sarah, AL Hughes, Kate Starbird, and Leysia Palen. 2010. “Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness.” *Proceedings of the 28th …*: 1079–1088. http://dl.acm.org/citation.cfm?id=1753486.

Wang, Tian. 2012. “Search for a New Perspective.” *Twitter Blog*, November 15. http://blog.twitter.com/2012/11/search-for-new-perspective.html.

Weng, Jianshu, Ee-Peng Lim, Qi He, and Cane Wing-Ki Leung. 2010. “What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter.” *2010 IEEE International Conference on Data Mining* (December): 1121–1126. doi:10.1109/ICDM.2010.34. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5694095.

Whiting, Stewart, I Klampanos, and J Jose. 2012. “Temporal Pseudo-relevance Feedback in Microblog Retrieval.” *Advances in Information Retrieval*: 522–526. http://www.springerlink.com/index/L4791425176141T3.pdf.

Craig Willis, Richard Medlin, and Jaime Arguello. Incorporating Temporal Information in Microblog Retrieval. In Proceedings of the 21st Text REtrieval Conference (TREC'12), National Institute of Standards and Technology, special publication, 2012.