Beyond Twitter Text: A preliminary Study on Twitter Hyperlink and its Application

Dehong Gao, Wenjie Li, Renxian Zhang
Department of Computing, the Hong Kong Polytechnic University. Hong Kong
{csdgao, cswjli, csrzhang}@comp.polyu.edu.hk

ABSTRACT

While the popularity of Twitter brings a plethora of Twitter researches, short, plain and informal tweet texts limit the research progress. This paper aims to investigate whether hyperlinks in tweets and their linked pages can be used to discover rich information for Twitter applications. The statistical analysis on the analysed hyperlinks offers the evidence that tweets contain a large amount of hyperlinks and a high percentage of hyperlinks introduce substantial and informative information from external resources. The usage of hyperlinks is examined on a self-defined hyperlink recommendation task. The recommended hyperlinks can not only provide more descriptive or explanatory information for the corresponding trending topics, but also pave the way for further applications, such as Twitter summarization.

KEYWORDS : Twitter, Hyperlink Usage, Hyperlink Recommendation

1. Introduction

The shift of information center from the mainstream media to the general public drives a growth of social network sites among which Twitter is undoubtedly one of the popular applications now. In academia, Twitter researches have become a new hotspot (H. Kwak et al, 2010; D. Zhao and M. Rosson, 2009; M. Michael and N. Kouda, 2010). Existing researches mainly focus on tweet content and user communication, while ignoring external resources. However, the limitation is the plain and short tweet text, which contains 140 characters at most. Even worse, tweets are usually written with many informal expressions.

It has been reported in (TechInfo, 2010) that about 25% of tweets contain hyperlinks and the proportion is increasing. Recently, even Twitter itself also provides an interface to allow people pay close attention to the tweets with hyperlinks. This move probably implies that (1) tweets contain a large amount of hyperlinks; and (2) hyperlinks in tweets may provide useful information for understanding topics. For instance, at the time when we write this paper, "*William & Kate*" is a popular topic of conversation. When we follow some arbitrary hyperlinks in the tweets under this topic, we are often directed to the Web pages containing the detailed information about their royal honeymoon (http://styleite.com/gjhao), the video of royal wedding (http://bit.ly/ld72jW) and the comments on their expensive wedding (http://ow.ly/1cKi99). Without length limit, a Web page tends to use a longer text describing the topics. Especially some of these hyperlinked pages are written by professional editors, and are much more regular than ordinary tweets. "*William & Kate*" is just an example here. But it motivates us to seek for the answers for the following questions:

Q1: How popular are hyperlinks in tweets?
Q2: Can these hyperlinks provide additional, useful and relevant information for understanding topics?

Proceedings of COLING 2012: Demonstration Papers, pages 155–162,
COLING 2012, Mumbai, December 2012.

Q3: What kinds of information can be explored from hyperlinked Web pages?

To answer these questions, we download ten trending topics from Twitter.com and annotate 2018 hyperlinks in selected tweets to categorize the information presented in the hyperlinked pages. The statistical analysis of the annotation results indicates that 44% of tweets contain hyperlinks, among them 35% of examined hyperlinks are worth are worth further exploration, Actually, our study on hyperlink usage analysis indicates that about 70% of those valuable hyperlinked pages provide descriptive or explanatory information regarding the topics, which is much richer and more formal than that brought by tweets themselves. All these statistics suggests that hyperlinked pages can be used as external resources to assist understanding or interpretation of topics. They have many potential uses in Twitter applications. For example, *trending topics explanation and summarization*, can be generated from hyperlinked pages, rather than just from obscure and incoherent summary with informal tweets.

In this paper, the usage of hyperlinks is examined on a self-defined hyperlink recommendation task. Here, *hyperlink recommendation* is defined to recommend the high-quality hyperlinks (i.e., the hyperlinks that provide most relevant textual information beyond Twitter text) for trending topics. The task is cast as a typical binary classification problem. Considering the small size of available labelled dataset and unlimited unlabelled dataset, a semi-supervised learning technique, called co-training, is employed to Leveraging the huge amount of unlabelled data resource to enhance classification performance. The results are promising. We hope our study will shed some light on the researches of Twitter hyperlink and its applications.

2. Hyperlink Analysis

To collect enough tweets, the public Twitter APIs (dev.twitter.com) are used to download tweets from the Twitter websites in real-time. We manually select ten trending topics from Twitter.com and download 81,530 related tweets with Twitter APIs from March 2 to March 12, 2011. These trending topics cover the main categories of trending topics. To answer the question 1 in Introduction, analysis is carried out on these tweets to determine the ratio of tweets with hyperlinks. In overall, the tweets with hyperlinks account for 44% of all the tweets, which is much higher than the data (25%) reported in (TechInfo, 2010) in July, 2010. This statistics answers the first question raised in Section 1. Meanwhile, we can also observe that the tweets belonging to the technology and emergency categories are more likely to include hyperlinks.

With such a high proportion of tweets containing hyperlinks, the next issue concerned is to examine the percentage of the hyperlinks that can provide additional relevant information about the topic, or to say how many of the posted hyperlinks are useful for understanding or interpreting the topic (and the tweets about the topic`). This is to answer the question 2 in Introduction. To this end, we randomly selected 2018 hyperlinks for further analysis. These hyperlinks are annotated as *useful, off-topic, spam, error-links,* and *non-English* hyperlinks. The term *useful* here means that the Web pages can provide relevant information for trending topics while the *spam* hyperlinks refer to the Web pages with the evidential intention of advertising, e.g. advertisements, or some e-commercial pages. The *off-topic* hyperlinks are those pointing to the Web pages with no relevant information and the *error-link* hyperlinks are invalid ones. As exhibited in Figure 1.(a), among 2018 selected hyperlinks, the useful hyperlinks take up 35%. A much higher ratio is shown in the technology and emergency categories, while the ratio of meme category is much lower than the others. These observations indicate that the amount of hyperlinks in tweets can be used to provide related information for given trending topics. Meanwhile, the higher ratio in technology and emergency and the lower in meme category also indicate that the

divergence of useful hyperlinks across the categories is striking. We also find that most non-useful hyperlinks are off-topic hyperlinks (42%), especially in Meme category. This indicates the necessity of useful hyperlinks identification.

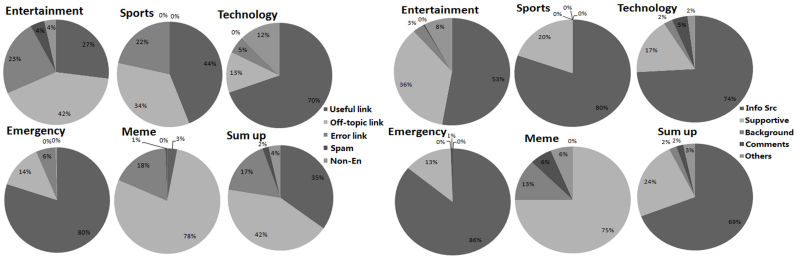

FIGURE 1 – (a) Proportion of useful hyperlinks; (b) Proportion of hyperlink purpose

To answer the question 3, the purpose of useful hyperlinked pages is investigated to evaluate what kind of information can be discovered from hyperlinked pages. We categorize the purpose of useful hyperlinked pages as *source information*, *background information*, *supportive information*, *comment* and *others*. These categories can direct researchers to explore different applications, e.g. information source Web pages is kind of help in trending topic explanations and summarization while the comment Web pages tends to be important in opinion mining. As Figure 1.(b) suggests 69% of the annotated useful hyperlinked pages convey information about the source of trending topics, which can be also regarded as explanations to the trending topics. Similarly, the higher ratios occur in the technology and emergency trending category and a lower ratio appears in the meme category.

Now we can conclude that tweets contain a large amount of hyperlinks and a high percentage of hyperlinks introduce substantial and informative information from external resources, though the quality of hyperlinks vary from category to category.

3. Hyperlink Recommendation

Since the statistics analysis on Twitter hyperlinks suggests that it is worth to explore the external information, we propose a new task, namely *hyperlink recommendation*, to recommend useful hyperlinks that provide most relevant information that beyond Twitter text for the corresponding trending topics. The significance of this hyperlink recommendation task is to pave the way for future researches that may need to identify the useful hyperlinks beforehand in order to enrich the text presentation of tweets or discover topic background information etc.

In this preliminary study, hyperlink recommendation is cast as a typical binary classification problem by separating the useful hyperlinks that provide the relevance information for a trending topic from the others. Useful hyperlinks are further ranked according to their relevance and credibility. The top ones are regards as the recommended hyperlinks. One problem is the insufficient labelled data for effective learning. Fortunately, compared with the limited annotated hyperlinks, a large amount of unlabelled Twitter hyperlinks are available from Twitter.com. Thus, we choose to adopt semi-supervised learning as a solution to incorporate the unlabelled data to enhance the classification performance.

We consider two sets of features for this classification task. They are the features related to tweets and topics of tweets, such as trending topics category and the average tweet length, and the features related to the hyperlinks and the linked target pages, such as PageRank (PR) value of the page and the domain of links (see Table 1). While a co-training technique is devised based on the assumptions that the given task can be described with two redundant feature sets and each set is sufficient enough for classification (M. Li and Z. Zhou, 2007), the design of these two independent sets of features is just coincident with the co-training assumption.

Features	Set	Examples
Trending Topics Category	Twt	*Entertainment, Meme, etc.*
Trending Created Time	Twt	*ipad2"created_at":"2011-05-19T00:44:13", etc.*
Spelling Error	Twt	*"frinds-> friends", etc.*
Acronym / Emoticon	Twt	*FAQ, 10x, Lemeno, etc./ "☺", "☹", "^o^"*
Repeating letter word	Twt	*"soooo","Awesome!!!!" , etc.*
Average (tweet length)/(sentence number)	Twt	*(12, etc.)/(2, etc.)*
Similarity of tweet and trending	Twt	*0.17817, etc.*
PR value	Hyl	*"www.apple.com" (PR value:9)*
Domain feature	Hyl	*cnn or yahoo*
HF-ITF	Hyl	*0.4037, etc.*
Similarity of webpage and trending/tweets	Hyl	*0.1085, etc.*

TABLE 1 – Hyperlink features and examples, where Twt/Hyl denote tweet/hyperlink feature set

As mentioned in Section 2, the quality of hyperlinks in different trending topic categories varies distinctly. Thus, the trending topic category is selected as one of the features. The created time of a trending topic is introduced as well. These two topic features can be extracted via the public APIs of "What the Trend" (www.whatthetrend.com). Regarding the tweet feature extraction, two perspectives are considered: text writing style and statistic information. The text writing style features focus on text expressions, including spelling error, acronym usage, emoticon usage, repeating letter word usage. The toolkits of Jazzy API are used to detect the spelling error. However, identifying the real spelling error in tweet is a challenging task when Twitter frequently broadcasts the messages that contains a large number of nouns not presents in a common dictionary (M. Kaufmann, 2010). The spelling checker will fail to recognize the correct nouns. Hence, we calculate the edit distance of the erroneous word and the suggested spelling by Jazzy. Only those pairs with distance less than 3 letters are corrected, like "*frinds-> friends*", "*Niether-> Neither*". For the acronym feature, we uses two publicly available acronyms word lists from SmartDefine (www.smartdefine.org) (e.g., "*FAQ: Frequently Asked Questions*") and ChatOnline (www.chatslang.com) (e.g., "*10x: Thanks*"). Meanwhile, ChatOnline offers 273 commonly-used emoticons like "☺", "☹" and "^o^". These emoticons are used to detect the emoticon usage in tweets text. The last feature about tweet writing style is the use of words with repeating letter or punctuation, like "*what'sssss up*", "*soooo guilty*" and "*Awesome!!!!*" etc. As for the tweet statistics features, average tweet length, average sentence words in tweets are concerned, and these features in certain degree reflect how much attention the tweet poster pay to the trending topic. Intuitively more attention implies that the tweet is more likely to contain a useful hyperlink. In case of hyperlink per se features, the domain information and Google Page-rank value are extracted, which indicate the global importance of hyperlinks. *hf-itf* of a hyperlink is also introduced, where *hf* denotes the frequency of hyperlinks present in one topic, and *itf* denotes the frequency of hyperlinks across all the topics. We also calculate the cosine similarity

between a hyperlinked page text and the collective text of the tweets containing that hyperlink, and between a hyperlinked pages and the trending topic it resides to indicate the content similarity.

Typical co-training paradigm works as follows. Given a set of labelled instances L and a set of unlabelled instances U, co-training will first define two features views (V_1 and V_2, corresponding to the feature sets related to tweets/topics and related to hyperlinks in this study) on each instance in L and U, and specialize two classifiers (C_1 and C_2) on each view. In each iterative procedure, classifier C_1 and C_2 will be trained with labelled instance L on feature set V_1 and V_2 respectively and label all instance in U. Then, the n most confident new labelled instances L' are allowed to update labelled dataset of L for the next iteration. The iteration terminates when all unlabelled instance U are labelled or nothing is changed. Normally, when parameter n increases, the quality of new labelled instances will decline and lead the unstable of co-training. To avoid this problem, we define the most confident instances as the ones with the same predicted labels by both classifiers. The advantage of co-training is that with sufficient feature sets, classifier C_1 and classifier C_2 can learn information from the unlabelled dataset and exchange it with the other one.

Given the identified useful links for a given topic, we further rank them in terms of relevance and credibility and recommend the top ones to users. The ranking strategy is simple. We first rank the useful hyperlinks by their PR values which represent the global importance of hyperlinks. When several links receive the equal PR value, they are ordered according to the similarity between the hyperlinked page and the trending topic.

4. Experiments and Discussion

The experiments are conducted on the whole hyperlinks extracted from tweets, including both labeled and unlabeled hyperlinks. We come up with 230 labeled hyperlinks. We are randomly split into the training and test datasets (with ratio 4:1). Similarly, unlabeled hyperlinks are generated from the unlabeled hyperlinks. The evaluations of supervised learning on different feature sets are provided in Table 2 as the baseline of comparison. Figure 2 shows the co-training evaluations using different number of new labeled instances updated in each iteration (n). Comparing Table 2 and Figure 2, the significant improvement is observed in the precision of co-training, which indicates the co-training can utilize two feature sets information to predict higher accuracy labels. The higher precision and lower recall of NB classifier reflect that co-training can help NB learn some accurate rules to precisely predict the label, but cannot balance with the coverage. In contrast, SVM is able to learn some complex "rules", which can cover more instances, while the precision declines a bit.

	NB P	NB R	NB F	SVM P	SVM R	SVM F
Integration	0.35	0.91	0.51	0.78	0.91	0.84
Tweet	0.64	0.95	**0.76**	0.85	0.85	**0.85**
Hyperlink	0.35	0.91	0.51	0.35	0.91	0.51

TABLE 2 – Evaluation of supervised learning

Additionally, compared with the performance of supervised learning with hyperlink features, a remarkable increase can be achieved by introducing tweet features, which indicates tweet features play an important role in useful hyperlink classification. To certain degree, these tweet features can be regarded as features of tweet relevance, which means tweet relevance can be a good indicator of usefulness hyperlink. When taking a closer look at the wrongly predicted instances, we found two main sources of errors. The precision error mainly results from the non-text context type of the hyperlinked pages. For example, a video page can be regarded as useful by manual annotation, but with little text information it is hard for classifiers to predict correctly.

The recall error is mainly caused by the Web page that simply has word overlaps with a trending topic but not really related to it. For example, the Web page of selling ipad2 is prone to be regarded as related to the trending topic of the launch of apple's ipad2 by classifiers, but actually the page is not what the users care about.

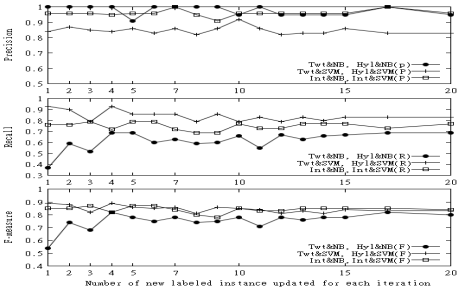

FIGURE 2 – Co-training evaluation

Eventually, useful hyperlinks are ranked by their PR value and their similarities with trending topics. The top two hyperlinks that are recommended for some example topics are illustrated in Table 3. These recommended hyperlinks provide the information about the trending topics, and help to understand the trending topics especially for technology and emergency category. For example, one of the recommended hyperlinks for "*#ipad2*" is the homepage of the Apple official website, and the other one presents the new properties of ipad2. The recommended hyperlinks for "Frankfurt Airport" are also linked to some predominate websites, e.g. jiHadwatch.com, webpartner.com, etc. However, the quality of recommended hyperlinks for meme trending topics is lower than the others. The reasons are: (1) the proportion of useful hyperlinks in meme is lower than that in the others and (2) there is too much unrelated information in this category. This also echoes the annotation findings.

In the future, we will continue to improve the performance of useful hyperlink classification by reducing the precision errors and recall errors. We would also like to further explore the usage of hyperlinks, and apply useful hyperlinks for potential applications.

Acknowledgements

The work presented in this paper is supported by a Hong Kong RGC project (No. PolyU 5230/08E).

Trending	**Recommended Hyperlink**
Frankfurt Airport	1. http://bit.ly/g9hcTN (jiHadwatch.com) 2. http://bit.ly/hymUEl (webpartner.com)
#ipad2	1. http://www.apple.com/ (apple.com) 2. http://on.mash.to/dYhMHa (mashable.com)
#ilovemyfans	1. http://bit.ly/iPTTon (twitpic.com) 2. http://mysp.ac/dH6vla (myspace.com)
Jon Diebler	1. http://bit.ly/lZ64qe (Yahoo.com) 2. http://bit.ly/gZrLa2 (buzztap.com)
Adonis DNA	1. http://bit.ly/e7TgOv (personalinjuryattorneyz.us) 2. http://aol.it/ghsgND (popeater.com)

TABLE 3 – Examples of recommended hyperlinks and their domains

References

H. Kwak and C. Lee et al, 2010. *What is Twitter, a social Network or a news Media*? ACM WWW10. Raleigh, North Carolina, USA.

D. Zhao and M. Rosson. (2009). *How and Why People Twitter: The Role that micro-blogging plays in informal communication at work*. ACM, GROUP09. Sanibel Island, Florida, USA.

M. Michael and N. Koudas. (2010). *TwitterMonitor: Trend Detection over the twitter stream*. ACM SIGMOD10. Indiana, USA.

TechInfo, 2010. http://bit.ly/muk2Uu.

M. Li and Z. Zhou. (2007). *Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples*. IEEE Transactions on Systems, Man and Cybernetics.

M. Kaufmann. (2010). *Syntactic Normalization of Twitter Message*. ICON10, IIT Kharagpur, India.