Investigating the Statistical Properties of User-Generated Documents

Giacomo Inches[1], Mark J. Carman[2], and Fabio Crestani[1]

[1] Faculty of Informatics, University of Lugano, Lugano, Switzerland
{giacomo.inches, fabio.crestani}@usi.ch
[2] Faculty of Information Technology, Monash University, Melbourne, Australia
mark.carman@monash.edu

Abstract. The importance of the Internet as a communication medium is reflected in the large amount of documents being generated every day by users of the different services that take place online. In this work we aim at analyzing the properties of these online user-generated documents for some of the established services over the Internet (Kongregate, Twitter, Myspace and Slashdot) and comparing them with a consolidated collection of standard information retrieval documents (from the Wall Street Journal, Associated Press and Financial Times, as part of the TREC ad-hoc collection). We investigate features such as document similarity, term burstiness, emoticons and Part-Of-Speech analysis, highlighting the applicability and limits of traditional content analysis and indexing techniques used in information retrieval to the new online user-generated documents.

1 Introduction and Motivations

Communication is a primary need of the human being and the advent of the Internet amplified the possibilities of communication of individuals and the masses [1]. As result, many people every day use chat and instant messaging programs to get in touch with friends or family, or rely on online services such as blog or social networks to share their emotions and thoughts with the Internet community [2].

The increasing popularity of these online-based services (Twitter, Facebook, IRC, Myspace, blogs, just to mention few of them) results in a production of a huge number of documents generated by Internet users. It is therefore of great interest to study the properties of these online user-generated documents: from a commercial point of view we could identify new trends and hot topics by mining them [3, 4], so as better focus advertisement or new online services; from a policing perspective, instead, it may allow us to detect misbehaviour [5–7]. Again, it is also interesting from a research point of view to understand the linguistic properties of such documents or their statistical properties to improve the current models and techniques used in Information Retrieval [8]. Since this kind of documents are more recent and less studied, compared to more consolidated collections (like e.g. the TREC ad hoc), we need to understand them as clearly

as possible in order to perform effective and valuable mining and/or retrieval on them [9].

We discuss the motivating related work in Section 2 and present the datasets used for our analysis in Section 3. In Section 4 we describe the analysis performed and the metrics used and conclude in Section 5 with a summary of the results of our study and an overview on the future work.

2 Related Work

Inches at al. [10] provided a preliminary analysis of the statistical properties of online user-generated documents. The authors show their "shortness" in terms of average document length and their "messy" nature due to spelling/mistyping errors as well as the fact that the terms occurrences followed a standard Zipfian distribution. Other works on online short documents focused more on clustering [11], on topic detection [12] or similarity measures [13], but without considering the general properties of the different collections analysed in each work.

Some work has already been done in trying to categorize the online documents based on their properties [7], leading to a distinction between chat- and discussion-like documents. More general properties of text can be found in the work of Serrano et al. [14], where the authors focus on different properties of "standard" written text with the purpose of developing a new and more complete model for the description of written text. More general purpose introductions to textual analysis can be found in [15] and [16].

This work aims at extending the analysis in [10] with the study of standard text properties such as the ones presented in [14] and to integrate them with a Part-Of-Speech (POS) study. To the best of our knowledge, neither of these analysis has been thus far performed on such collections. Instead, POS analysis has already been applied to queries [17], term weighting [18] and on text blocks [19], just to mention some topics.

3 Datasets

Our analysis aims at comparing user-generated documents and standard information retrieval ones, therefore we choose as representative of the first class four datasets containing different user-generated content, namely *Kongregate* (Internet Relay Chat of online gamers), *Twitter* (short messages), *Myspace* (forum discussions) and *Slashdot* (comments on news-posts). These datasets were first presented at the Workshop for Content Analysis in Web 2.0 [20] and are divided between training and testing data[3]. Our analysis take into consideration only the train dataset for each collection, which is enough to our purposes.

As collections representative of standard information retrieval documents we employed three datasets of similar edited content: news articles from the *Associated Press* (AP, all years), the *Financial Times Limited* (FT, all years) and

[3] Datasets and details available at `http://caw2.barcelonamedia.org/`

the *Wall Street Journal* (WSJ, all years). These datasets form a representative subset of the standard TREC Ad-hoc collection[4] and, although they are similar in the type of content, they cover different topics: *AP* and *WSJ* report news in general, while *FT* focuses on markets and finance.

We notice that these collections show a similar topicality to the particular *Myspace* and *Slashdot* datasets we use: The *Myspace* dataset covers the themes of campus life, news & politics and movies, while the *Slashdot* dataset is limited to discussions of politics. The fact that the themes are similar to the news articles is important in order to make statistical comparison between the collections meaningful. As for the topicality of the *Twitter* and *Kongregate* datasets, due to their conversational and more unpredictable nature, we cannot state precisely what their topicality is [21, 3, 12].

We report in Table 1 some basic statistics about these datasets. The difference in the average document length is evident: the user-generated document collections contain documents that are remarkably short compared to the news articles. We will examine in Section 4 the implications of this property in terms of the document self-similarity and burstiness, where we will explain also the role of common and rare words.

Table 1. Statistics of datasets

	avg. doc. length (# words)	avg. word length	# Common words (% of the vocabulary)	# Rare words (% of the vocabulary)
Kongregate	4.50	7.55	489 (1.39)	29'805 (84.65)
Twitter	13.90	7.30	716 (0.20)	354'131 (97.19)
Myspace	38.08	8.11	743 (0.39)	179'757 (96.10)
Slashdot	98.91	7.88	560 (0.45)	118'276 (95.88)
WSJ	452.00	7.57	1003 (0.44)	219'332 (96.85)
AP	464.23	7.53	1217 (0.40)	298944 (97.34)
FT	401.22	7.26	1017 (0.36)	271'055 (97.23)

4 Analysis of the Datasets

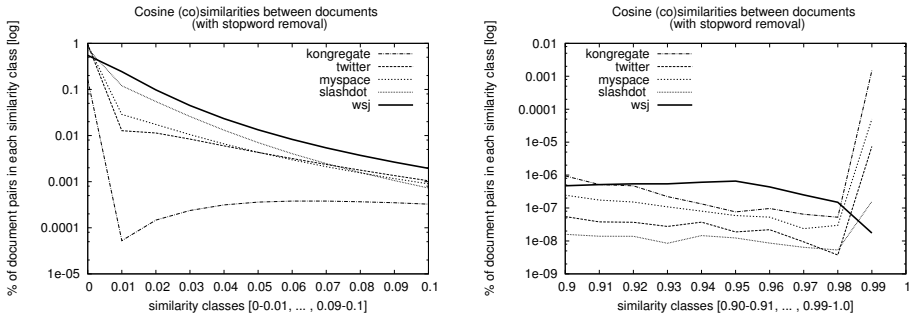
4.1 Similarity

The first property that we study is the self-similarity between documents, which we compute using the cosine similarity between td-idf document vectors.

The comparison was performed between each pair of documents in the collection for a total of $\frac{N(N-1)}{2}$ comparisons for each collection (where N is the number of documents in the collection, available in [20, 10]). We choose the WSJ to represent the TREC collections and display the values of the self-similarity

[4] Datasets and details available at `http://trec.nist.gov/data/test_coll.html`

computed after having removed the stopwords[5] from the documents. The most evident difference between the user-generated documents (*Kongregate*, *Twitter*, *Myspace* and *Slashdot*) and the standard ones (represented by the *WSJ*) can be observed at the extremes of the similarity scale. For this reason, in Fig. 1 we zoom in to show only the percentage of document pairs with the lowest (left) and highest (right) similarity scores.



(a) First similarity classes (b) Last similarity classes

Fig. 1. Self-similarity between documents after stopword removal. We normalized the count for document in each similarity class by the total number of comparisons.

In the first case, we observe that user-generated documents appear less frequently with lower similarity values (0.01-0.09), as they become shorter. To the contrary, they appear more frequently with higher similarity values (0.9-1.00), contrasting the behaviour of the standard documents. The latter, in fact, drop down when we consider only the last similarity range (0.99-1.00).

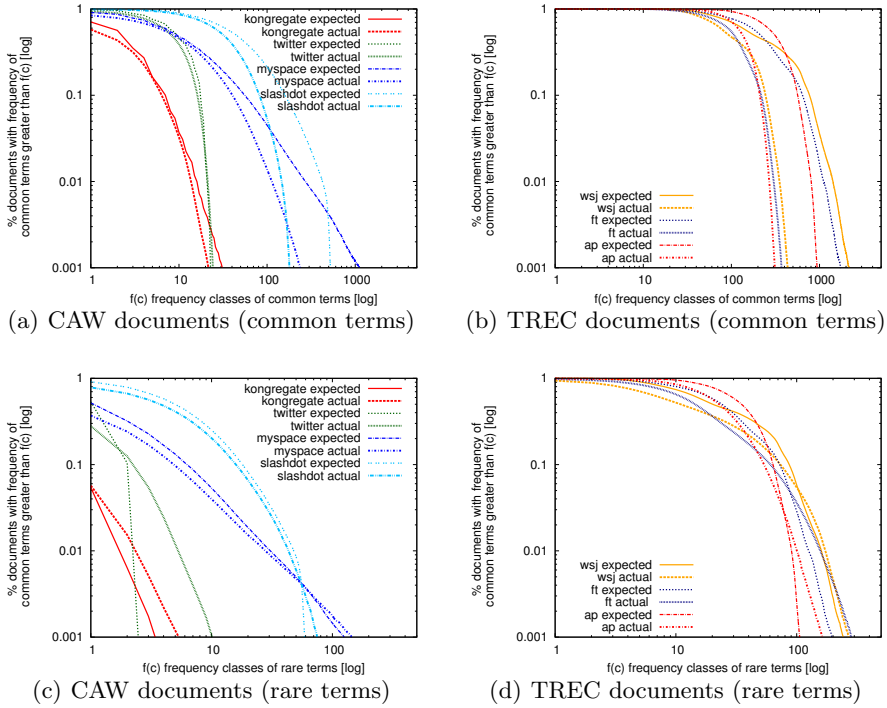
This means that shorter documents seem to be more similar across themselves than the longer ones. This can be explained with the length of the documents itself: short documents contain less words (less "information"). Therefore, given two short documents, there is an higher probability that they appear to be similar even if they are unrelated, just because they are short.

To counteract this behaviour of the shorter user-generated documents we would need to enlarge the information they carry whenever we want to process them. Different solutions can be applied to this problem, which we leave for future study. We just mention two techniques which we could use: stream segmentation and document expansion. With stream segmentation we aim at merging documents together based on their temporal proximity, which raises the problem of setting proper boundaries for joining them, while with document expansion we could extract relevant information from the documents, such as internet links or tags, and retrieve from other sources new words to enlarge their topicality.

[5] Standard Terrier stopwords list

4.2 Burstiness

We present in this section the second analysis, where we study the burstiness of the terms in each collection. Plots in Fig. 2 show the percentage of documents in each collection that contains a certain number of common or rare words. Common words are defined as the most frequent words in the vocabulary that account for more than 71% of the text in the collection, while rare words are the least frequent words in the vocabulary that account for 8% of the text, as computed also in [14] (Table 1).



(a) CAW documents (common terms)

(b) TREC documents (common terms)

(c) CAW documents (rare terms)

(d) TREC documents (rare terms)

Fig. 2. Common and rare term burstiness for user-generated documents (CAW) and traditional ones (TREC).

In each plot we show also the expected number of such documents if the words in the vocabulary were uniformly distributed (according to their overall frequency in the collection) across the documents in the collection. Differences between the curves for actual and expected number of documents indicates burstiness in the collection, i.e. the phenomenon that a word observed once within a document is far more likely to re-occur within the same document than it is to occur in another document chosen at random.

Looking at the common terms plot for the three edited collections (*AP*,*FT* and *WSJ*), we see that the line denoting the actual number of documents with a certain number of common terms in them lies well below the expected number of such documents. This indicates that documents are bursty, since common terms are not spread evenly across the collection of documents, but are concentrated more in some documents than others. The same is true (although to a less extent) for the rare terms in these collections: the actual number of documents containing a certain number of rare words lies below the expected curve, again indicating that documents are bursty, since the rare words are not uniformly distributed across documents.

Comparing the plots for user-generated content (*Kongregate*, *Twitter*, *Myspace* and *Slashdot*) with those for the edited collections, we see that the difference between the expected and actual number of documents is far less pronounced for the new collection (especially for the common terms) than it is for the traditional ones. This indicates that burstiness may not be an important issue for user-generated content as it is for traditional collections. This may have implications in document-length normalization for these collections: as we already noticed in Section 4.1 we should eventually pre-process them and expand their informative content through the use of document expansion or segmentation.

The fact that the expected/actual curves for the different user-generated collections differ greatly from one another is due to the large difference in average document length in the different collections. The curves for the edited collections (especially for the common terms) line up quite well due to the fact that the average document length is very similar.

4.3 Part-Of-Speech Distribution

In the third part of our work we employ GATE[6] and its built-in tokenizer, sentence splitter and Part-Of-Speech (POS) analyser called ANNIE[7] [22, 23] to analyse the Part-Of-Speech (POS) tags distribution in the different datasets.

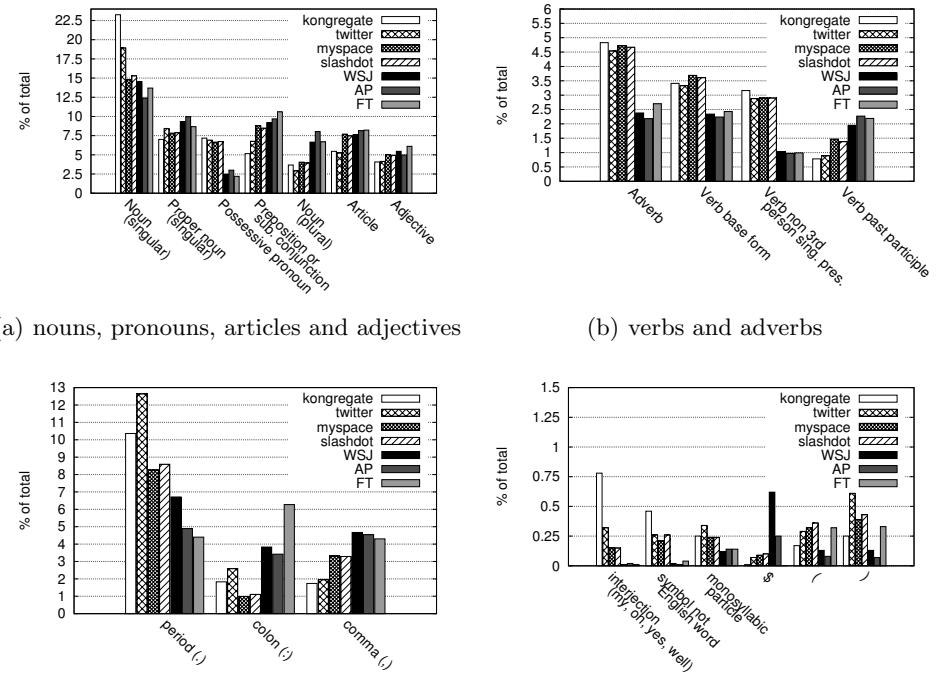
We report in Fig. 3 the results of the POS extraction through ANNIE of the full text on 30% of the documents in the collection, selected at random (since we did not find significant variation in the distributions with an higher subset). We used the ANNIE default settings for each component of the processing chain (tokenizer, sentence splitter and POS extractor) and report in Fig. 3 only the most significant categories[8].

If we study in detail the results of Fig. 3 we can observe two different collection behaviors: first, we notice some inter-collection variations, between the user-generated datasets and the traditional datasets, then we perceive an intra-collection variation, inside the user-generated datasets, between chat-style and discussion-style documents.

[6] GATE: "General Architecture for Text Engineering", `http://gate.ac.uk/`

[7] ANNIE: "A Nearly-New Information Extraction System", `http://gate.ac.uk/`

[8] A complete list of the POS tag extracted by ANNIE can be found on `http://tinyurl.com/gate-pos`



(a) nouns, pronouns, articles and adjectives

(b) verbs and adverbs

(c) high frequency punctuation

(d) interjections, symbols and low frequency punctuation

Fig. 3. POS analysis

Inter-collection differences can be seen in the usage of proper nouns, possessive pronouns and plural noun Fig. 3(a) as well as in the usage of verb and adverb Fig. 3(b). An explanation for this may be found in the nature of the documents contained in each collection: in the user-generated texts the user producing them is focused in expressing his/her point of view or emotions against the others (high usage of possessive pronouns), qualifying the amount of their sensations (high usage of adverb), addressing directly in first person (high usage of verb not in the third person singular) and referring to action occurring mostly in the present time (verb in base form). To the contrary, texts that are edited in a professional way report events occurred in the past (high usage of verb in past participle), not occurring to the author itself (high usage of third person in the verb) or taking place in a particular location (higher use of singular proper noun). Again, if we take a detailed look at the punctuation, interjection and symbols in Fig. 3(c) and Fig. 3(d) we observe how user-generated documents consist of a more direct, personal and simple communication, given by a more extensive usage of interjection, symbols, monosyllabic particles and periods. Edited content, instead,

is more descriptive, due to the usage of colons and commas, which generally link together different concepts inside the same sentence. A last observation regards the usage of brackets, which are more employed in the user-generated documents. We suppose they should be used in combination with colons and semicolons, building the so called emoticons, to enrich the expressiveness of the communication. We therefore analyse the usage of the emoticons in Section 4.4.

Intra-collection differences can be seen within the user-generated collection, where some datasets (*Myspace* and *Slashdot*) appear to be more related to the edited texts than the others (*Kongregate*, *Myspace*), which highlight different properties. These properties are an high usage of proper singular nouns, periods, interjections and symbol, and a less usage of articles and adjectives, which becomes the least among all the collection for verbs in the past form and commas. They can be seen as attributes of an essential and immediate communication, such as the online-chat (*Kongregate*) or similar to chat (*Twitter*). On the other hand, for some POS categories the *Myspace* and *Slashdot* datasets are similar or just in-between to and with the trec datasets: this appear for preposition and subordinative conjunction, adjectives (Fig. 3(a)), verb in the past partiple form (Fig. 3(b)) as well as for periods, commas (Fig. 3(c)) and interjections (Fig. 3(d)). Following the approach proposed in [7], we label these documents as discussion-style documents.

These inter-collection and intra-collection differences can be used together with the measure of similarity and burstiness to give a preliminary classification of a dataset of unseen documents (standard/edited content or user-generate content, if user generated, chat- or discussion-style) as well as to help the retrieval of documents from a collection of a given type.

4.4 Emoticons and "Shoutings" Distribution

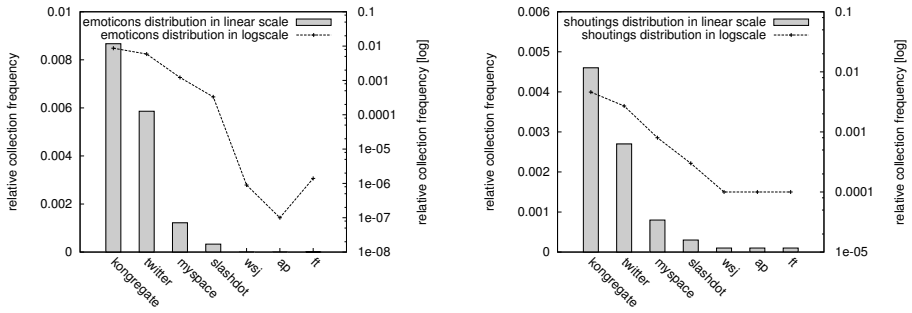
In this last part of our work we complement the POS analysis of Section 4.3 by investigating the distribution of emoticons and "shoutings" among the different collections. These features, in fact, can be very discriminative for identifying user-generated content [24] and in particular conversational data [3].

We collected a list of the most common emoticons (mostly through Wikipedia, see attachment A for a complete list) and parsed each document by comparing each token separately with a regular expression, thus identifying and counting only whitespace separated emoticons (such as `:)` and `:P`).[9] In a similar way we counted so-called "shoutings", that we define as whitespace separated tokens containing a succession of three-or-more consecutive instances of the same letter (e.g. `zzzz` and `mmmmaybe`). We did not include in this count tokens containing internet addresses (`www` and `WWW`) since they does not provide additional information on the collections being analysed.

In Fig. 4 we report the distribution of the emoticons and shoutings among the collections. The values represented are the relative collection frequency in both

[9] We experimented also with matching emoticons within sequences of characters like `hello:)mum` but obtained too many false positives to consider those results valid. For the same reason, we did not count emoticons containing whitespaces such as `:␣)`.

the linear and log scale. The behaviour of the distributions is similar and reflect the nature of the collections. User-generated collections (*Kongregate*, *Twitter*, *Myspace*, *Slashdot*) contain a large number of colloquial and informal tokens, such as emoticons and shoutings, that are used to improve the expressiveness of the communication. In the more standard and "professional" documents (*WSJ*, *AP*, *FT*), on the other hand, the communication remains on a formal and neutral level (having these collection almost zero counts for emoticons and shoutings and at least 1 order of magnitude less than the others).



(a) Emoticons distribution (b) Shoutings distribution

Fig. 4. Collection relative emoticons and shoutings distributions

As for the POS features analysed in Section 4.3, beside the inter-class differences highlighted above, it is also important to notice that we can highlight some intra-class differences among the user-generated documents: the more chat and colloquial documents (*Kongregate* and *Twitter*) contain more emoticons and shoutings occurrences (on the order of 1 or 2 levels of magnitude) than the documents that are more of a discussion-style.

These observations reinforce our conclusions about the usage of the features analysed in this work to improve the mining and retrieval of user-generated documents, which are nowadays of great interest for their novelty and popularity. To provide a practical example, we are currently applying these POS, emoticons and shoutings features to the TREC Blog08 collection[10][25] to improve the results for the Faceted Blog Distillation Task, to distinguish for example between personal (more colloquial) and formal (more neutral) blogs. Another application of these features, especially the emoticons, can be found in the problem of detecting opinionated blog content (where we can search for emoticons expressing particular feelings, as in Table 2).

[10] `http://trec.nist.gov/tracks.html`

Table 2. Top 10 emoticons in each dataset with their relative frequency as a percentage of all emoticon occurrences. We omit the few counts for WSJ,AP,FT since they are not informative. Emoticons in *italic* express a negative feeling (sadness), all the others a positive one (happiness, astonishment, smartness, tongue, smiley,...)

	Kongregate		Twitter		Myspace		Slashdot	
	emoticon	%	emoticon	%	emoticon	%	emoticon	%
1	:P	16.89	:)	43.35	:)	33.13	:)	37.26
2	XD	13.09	;)	11.12	;)	12.84	;)	17.75
3	:)	12.72	:-)	10.22	:P	10.84	:-)	14.92
4	:D	10.92	:D	8.78	:D	8.93	;-)	10.56
5	-.-	5.11	;-)	5.31	:]	4.61	:P	5.42
6	xD	4.62	:P	5.15	XD	3.47	:D	2.94
7	:O	3.45	*:-(*	1.82	:p	2.84	B)	1.94
8	=D	2.95	XD	1.42	=P	2.39	*:-(*	1.36
9	:p	2.84	:p	1.36	xD	2.37	:p	1.19
10	=P	2.72	:-D	1.10	:-)	1.61	:-P	1.04

5 Conclusions and Future Work

In this work we analysed two different collections, a new sets of user-generated documents and a traditional one, containing edited documents.

In the first part of our study we computed the self-similarity between pairs of documents and analysed the term burstiness in each collection. We were able to identify one issue related to the length of the documents: for user-generated documents, in fact, their extreme shortness (compared to the traditional ones) makes them to be too little informative (when they are a lot) or to much informative (when they are not). To this purpose, we identified two techniques which could be used to address this problem but we left their implementation and evaluation to future studies.

In the second part of our study we performed a Part-Of-Speech (POS) analysis on a representative samples of our collections and we found that there exist some significant differences in the usage of the grammatical elements within the different datasets: noun, verbs, adverb as well as punctuation, interjections and symbols can be used to distinguish between new user-generated and traditional edited-content. They can also be applied to identify more chat-style content than discussion-style text inside the class of user-generated documents, since the latter are more traditional content likely, while the first ones show up new properties.

In the last part of the study we reinforce the conclusions drawn in the second part of our work, noticing how two particular features, the emoticons and the shoutings, also allow us to identify differences between "traditional" and user-generated content. Moreover they allow to distinguish between different types of user-generated documents: chat-like and discussion-like.

Taking into consideration the results of all parts of this works, we could extend it focusing on the document expansion or stream segmentation for each

of the specific document classes: for chat-style and discussion-style documents we could consider the topic coherent portion of the entire chat or discussion as a document unit, instead of the single message or post (as we did in here). We are current applying the features discussed in this work to the TREC Blog Distillation Task and in future studies we plan to further refine the POS analysis, studying the occurrences of blocks of categories, similar to the work of Lioma and Ounis [19]. We also plan to tune the POS parser to better fit our collection, being able to detect categories (name-entity recognition) for the purpose of sentiment analysis [26, 27] or polarity detection. To conclude, we could also like to use the statistical properties identified in this study to classify the content type of an unknown collection (user-generated vs edited and/or chat vs. discussion) and use this information for resource selection.

A List of Emoticons Used

```
(Z.Z)  (-.-)Zzz  Zzz :) :-) :-]  :]  :] :> :-> => ^_^  ^-^  (^_^)  ^.^ ;) ;-) ;]  ;-] ;>  ;->
(^_~)  ^_~  ^_* :wink:  :(  :-(  :[  :-[ =(  =[  :< =<  :D  :-D  :D  :D  =D  X-D  XD
XD  xD  BD  8D X3  x3 :P  :-P  :-p  :p  =P  =p  :|  :-|  8)  8-)  B)  B-)  :'(  :'-( :'[
:'-[  ='(  ='[  :'<  :'-< ='< T_T  T.T (T_T)  Y_Y  Y.Y  (Y_Y) _  .  (_)  ;-;  ;_;  ;.;  :_:
._._.  :S  :-S  =S @_@  :-?  :?  ?_?  ?.?  :\")  :-\")  :/)  :-/)  :-O)  :O  :-o) :o :-O)  :O
=O  =O  =o  =_=  -.-  -.-\  -.-  o_o  o.o  o.O  O.o  O_o  o_O  -  (-)  _
(_)  (oO)  (Oo)  .  o_O  O_o  O.o  o.O  Oo  oO >_<  .  _  U_U  u_u  (U_U) <3  ()(_)()
U.U  u.u  (U.U)  U.u  V_V  v_v  (V_V)  V.V  v.v  (V.V)  <_<  >_> *_*  (*_*)  (*__*)  (*.*)
*.*  :-*  :*  :-x  :x  :-X  :X  ^*^ \\o/  _  (_)  x_x  X_X  (X_X)  x.x  X.X  (X.X)
#_# (#_#) OwO  (OwO) (*w*)  *w*  :-Q_  :-Q__ :-Q___  :Q_  :Q__  :Q___  : Q_  (:  (-:
: Q__  : Q___  :-Q :Q  =-Q_  =-Q__  =-Q___  =Q_  =Q__  =Q___  = Q_  = Q__  Y_Y  i.i  i_i
= Q___ =-Q  =Q  **  =3  :-3  :3  x3  :-@  :@  >:-@  >:@  >:  >_<  _  @>--
@:)  @:-)  @:-]  @:]  @:>  @:->  @=>  @=)  @=]  \\(^_^)/  *<:o)  ^o^ T.T  Y.Y  T_T
```

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, New York, NY, USA, ACM (2007) 56–65
2. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: WOSP '08: Proceedings of the first workshop on Online social networks, New York, NY, USA, ACM (2008) 19–24
3. Haichao Dong, S.C.H., He, Y.: Structural analysis of chat messages for topic detection. Online Information Review **30**(5) (2006) 496 – 516
4. Kucukyilmaz, T., Cambazoglu, B., Aykanat, C., Can, F.: Chat mining for gender prediction. Advances in Information Systems (2006) 274–283
5. Medina, E.W.: Military textual analysis and chat research. International Conference on Semantic Computing **0** (2008) 569–572
6. Bache, R., Crestani, F., Canter, D., Youngs, D.: Mining police digital archives to link criminal styles with offender characteristics. In: ICADL. (2007) 493–494
7. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain (2009)
8. Qi, H., Li, M., Gao, J., Li, S.: Information retrieval for short documents. Journal of Electronics (China) **23**(6) (2006) 933–936

9. Wang, F., Greer, J.: Retrieval of short documents from discussion forums. Advances in Artificial Intelligence (2002) 339–343
10. Inches, G., Carman, M., Crestani, F.: Statistics of online user-generated short documents. Advances in Information Retrieval (2010) 649–652
11. Carullo, M., Binaghi, E., Gallo, I.: An online document clustering technique for short web contents. Pattern Recognition Letters **30**(10) (2009) 870 – 876
12. Tuulos, V.H., Tirri, H.: Combining topic models and social networks for chat data mining. In: WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, IEEE Computer Society (2004) 206–213
13. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. Advances in Information Retrieval (2007) 16–27
14. Serrano, M., Flammini, A., Menczer, F.: Modeling statistical properties of written text. PLoS ONE **4**(4) (04 2009) e5372–
15. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
16. C.D.Manning, H.Schütze: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
17. Allan, J., Raghavan, H.: Using part-of-speech patterns to reduce query ambiguity. In: SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2002) 307–314
18. Lioma, C., Blanco, R.: Part of speech based term weighting for information retrieval. In: ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, Springer-Verlag (2009) 412–423
19. Lioma, C., Ounis, I.: Examining the content load of part of speech blocks for information retrieval. In: Proceedings of the COLING/ACL on Main conference poster sessions, Morristown, NJ, USA, Association for Computational Linguistics (2006) 531–538
20. J.Codina, A.Kaltenbrunner, J.Grivolla, E.Banchs, R., R.Baeza-Yates: Content analysis in web 2.0. In: 18th International World Wide Web Conference. (04 2009)
21. Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: ICWSM. (2010)
22. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)
23. Wilcock, G.: Introduction to linguistic annotation and text analytics. Synthesis Lectures on Human Language Technologies **2**(1) (2009) 1–159
24. Balog, K., Bron, M., He, J., Hofmann, K., Meij, E.J., de Rijke, M., Tsagkias, E., Weerkamp, W.: The university of amsterdam at trec 2009: Blog, web, entity, and relevance feedback. In: TREC 2009 Working Notes, NIST (November 2009)
25. Macdonald, C., Santos, R.L., Ounis, I., Soboroff, I.: Blog track research at trec. SIGIR Forum **44**(1) (2010) 58–75
26. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. (2010)
27. Ku, L.W., Ke, K.J., Chen, H.H.: Opinion analysis on caw 2.0 datasets. In: CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain (2009)