What's inside MySpace comments?

Luisa Massari
Dept. of Computer Science
University of Pavia
Pavia, Italy
email: luisa.massari@unipv.it

***Abstract*—The paper presents a characterization of the comments included in the user profiles of a popular social networking site. Some parameters are defined, which express comments composition in terms of length, language used, and external resources accessed. From the analysis of comments a model is derived, which reflects three different types of users. The behavior of users in terms of the language used is also derived.**

***Index Terms*—Social networks, comments, user profile characterization.**

I. Introduction

Social networks are online environments that let users to build a personal profile page, including personal details and cultural interests, to publish various types of content, such as, pictures, email, videos, blogs, comments, and to establish connections with other users to enlarge friend circle. Communication modes usually differ from face to face interpersonal relationships. Comments, in particular, represent for social network users a sort of one-to-many communication: comments are messages posted in personal pages, and made visible to everyone or to friends only. Comments can include text, references to Web pages, animated objects or images. Hence, comments represent a rich source of information about users opinions, connections, and also a way for easily and rapidly advertising and disseminating information and content.

The purpose of this study is to analyze and characterize the behavior of MySpace [1] users from the point of view of comments posted in their profile pages. MySpace has been chosen because of its popularity; even though Facebook leads social networking category, MySpace is still a top 10 website in the US. ComScore [2] reported 112 million unique US visitors for Facebook in December 2009, and 57 million for MySpace; 54% of all Internet users visited Facebook in December, and 27% visited MySpace [3]. Moreover, MySpace is one of the few social networking sites that allows to easily access to profile pages and retrieve data about users.

Social networks have recently captured the interest of different areas. Many companies analyze social networks in order to understand what web users, and people in general, think about their brands, their product and their services. For the purpose, companies rely upon search engines or a few analysis systems that mine social network public profile pages for specific keywords. In the research community, many studies have focused their attention on the analysis of online social networks, on their contents and on users characteristics. The nature, the structure, and the evolution of social networks are examined in many recent paper (see, e.g. [4]–[7]). Some studies ([6]) use graphs to describe social networks structural properties and to provide measures on the proximity between groups of users. The impact of social networking services is addressed in [8], where authors show that the majority of collaborations among users results from the opportunities of interactions offered by the services available on the sites. Some papers have addressed the characterization of the technological aspects of the workload of social networking web sites, in particular of sites offering specific services, such as, YouTube for video-sharing ([9]–[12]), Wikipedia for the creation of the so called wikis [13], and blogs [14]. These studies outline the peculiarities of these new types of workloads compared with the characteristics of traditional web workloads. A few recent papers address specifically the analysis of comments as a form of interpersonal communications. In [15] a large number of MySpace profiles are analyzed, and demographic characteristics are derived, together with a model of the language used. In [16], [17], the characteristics of social network comments are investigated. English comments are considered, and characteristics, such as length and language features are analyzed, together with dialogs between pairs of friends.

The analysis presented in this paper differs from previous work because of the large number of profiles considered and because, as far as we know, no study in the existing literature has approached the analysis of external resources accessed through comments and of the different languages used. In particular, our objective is to characterize the behavior of social networks users in terms of the type of language used in comments, of their length and of links to external resources. In Section II, we describe the data considered in this study and the methodology for their analysis. Section III presents results on user profile characterization, statistics on comments and on references to external resources. Moreover, the analysis of the words used in comments is presented. Finally, conclusions and future work are presented in Section IV.

II. Data Sample

A sample of 1.4 millions of public user profiles has been obtained by crawling, using the *curl* command line tool, the MySpace Web site [18]. These profiles contain about 369 millions of comments which have been analyzed to characterize the language used and the type of content published in MySpace profiles.

As a first step, external links have been identified in comments, in order to analyze locations of the object's data and Web resources. Remaining text has then been parsed and cleaned in order to remove special characters, HTML tags, punctuations, and to obtain a list of words. These words have been spell checked and valid words, that is, belonging to a dictionary, have been obtained. Most popular languages have been considered, namely, English, French, Italian, German, Spanish, Portuguese, and Dutch. Unrecognized words due to typos, or typical of the "Internet language" such as abbreviations, slang expressions exclamations have hence been identified.

At the end of this process, each comment has been described by its length in terms of number of words, valid and non-valid, and number of characters, by the language used, and by the number and types of external references. More in detail, each comment has been identified by two ids: the id of the corresponding and the id of the user who added the comment. These identifiers are assigned to users when they register. The characteristics of the comment in terms of language composition, length, and external resources have then been expressed by means of 16 parameters, namely, the number of raw and clean characters, that is, before and after cleaning, the total number of words, the number of words belonging to one of 7 languages (English, French, Italian, German, Spanish, Portuguese and Dutch), the number of unrecognized words, the number of links to Web pages, to images and to animated objects, the number of emoticons and of exclamations, which we have defined as long sequences of punctuation characters.

III. Statistics

Comments are distributed across the user profiles as shown in Figure 1. The figure represents 62% of the profiles, containing up to 100 comments. About 11% of the total, namely 149,185 users, have just one comment in their page, and 6% have two comments. 50% of the profiles have up to 40 comments.

A first raw characterization of comments has been obtained by some basic statistics. Table I shows, for some parameters, the maximum number of occurrences in one comment and the total number of occurrences over all comments. Minimum value of all parameters is always zero.

As can be seen, comments involve a huge amount of data, and justify a more detailed analysis of their characteristics.
By looking at the distribution of the previous parameters, it comes out that a large number of comments contains only text. Indeed, 89.3% of comments do not contain links, 86.3% do not contain images, and 99.2% do not contain objects. Looking at percentile values, we discover that 98% of comments have less than three external references and two images. Hence, it seems that external references are mainly concentrated in few comments. Moreover, 90% of the comments contain up to 42 words and up to 226 characters. 5.7% of the comments do not contain words.

In order to characterize user behavior, comment has been analyzed on a profile basis, that is, statistics on comments have been computed for each profile as average with respect

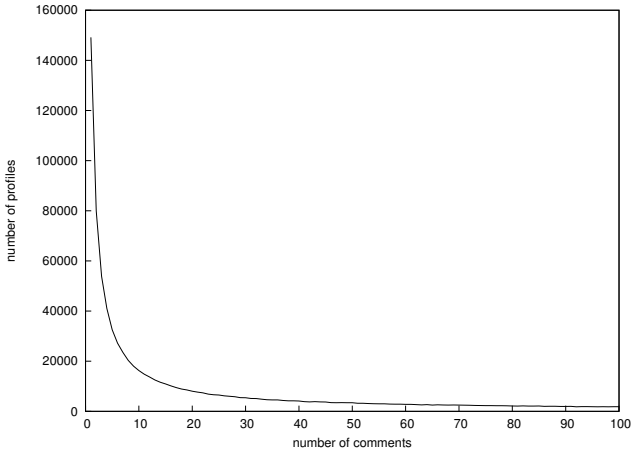

Fig. 1. Frequency distribution of comments.

	maximum	total
Web pages	563	47,876,519
images	1365	48,317,140
objects	22	2,771,047
words	138,024	7,905,129,273
clean characters	669,173	39,065,921,109
raw characters	681,270	63,086,834,409

TABLE I
Basic statistics for comments.

to the number of comments in the profile, hence obtaining a set of parameter values characterizing the single profile. Table II summarizes average values, standard deviation and quartiles values. Statistics for the number of objects are close to zero, and are not shown in the table. The minimum number of comments is one, while the minimum value for all other parameters is always zero.

Standard deviation values denote a high variability in parameter distribution. Moreover, looking at quartile values, we note that 3rd quartile is very close to the average, which, in turn, is always larger than the median, for all parameters. This means that all parameters distributions have long tails, that is, reveals the presence of parameters having values which are very high and different from the average behavior of the distribution. This characteristic will be better analyzed in the subsequent analysis.

Looking at the numbers of external references contained in a profiles, it comes out that on average, a profile contains 0.38 links in comments, 0.29 images, one exclamation and very few emoticons and objects. Moreover, 20% of user profiles, namely 290,000, have zero links in their comments.

In order to derive typical user behaviors, more advanced statistical techniques, such as clustering, have to be applied. Clustering is a multidimensional statistical analysis technique used to discover groups of points having similar characteristics. Clustering algorithms [19] partition a set of points into groups, or clusters, such that points belonging to the same cluster exhibit similar behavior, that is, distance among points within a cluster is smaller than the distance among points of

	average	std. dev.	1*st* quartile	2*nd* quartile	3*rd* quartile	maximum
comments	274.2	975.0	5.0	39.00	251.0	353,174.0
Web pages	0.3	0.6	0.02	0.16	0.5	40.0
images	0.2	0.4	0.04	0.17	0.4	61.0
clean characters	248.5	247.0	127.5	176.58	280.8	20,540.6
raw characters	132.0	170.7	69.3	95.6	131.2	13,986.0
total words	26.1	33.8	13.75	19.11	26.2	2,779.8
unrecognized words	2.5	5.8	1.0	1.49	2.5	1,108.0
valid words	23.6	31.8	12.02	17.35	23.7	2,507.0
English words	23.0	31.7	11.42	16.86	23.0	2,506.0
emoticons	0.1	0.6	0.00	0.12	0.2	522.5
exclamations	1.0	1.7	0.44	0.76	1.1	495.7

TABLE II
BASIC STATISTICS ON A PROFILE BASIS.

different groups. A cluster is represented by its centroid, that is, the geometric center of the group. From the point of view of clustering techniques, a user profile is represented as a point in a multidimensional space, the number of dimensions being the number of parameters used to characterize each profile. For the analysis of the user profiles we used the k-means clustering algorithm, in which the similarity criterion is based on the Euclidean distance. Since the objective of the study was to evaluate user characteristics related to comments, the number of comments and the number of links to Web pages and of total words contained in have been used as characterizing parameters. Other parameters have been discarded because highly correlated. Indeed, correlation indexes have been computed in order to discover dependencies. Highly correlated parameters are the number of English words, of valid words, of total words and of clean characters.
The *R* environment [20] has been used to compute correlations, to apply clustering, and to derive profile characteristics.

In cluster analysis, only profiles which are relevant from the point of view of content, having an average number of external references to Web pages larger that one, have been taken into account. Hence, 119,842 users have been considered. As a result of the clustering algorithm, users profile have been subdivided into three groups.

Table III summarizes centroids values and number of user profiles belonging to the three clusters. Cluster one groups 80.7% of user profiles, having on average 75.59 comments and 1.57 links. These represent the majority of users, having a high number of comments in their pages, and characterized by the smallest number of references to external links contained in. To the second cluster belong users having a high number of external references in their comment; these users correspond to the tail of the distribution of links. Then, there is a minority of users, about 7.5% of the considered profiles, belonging to group three, having a small number of comments in their pages, but have the highest number of total words. Moreover, looking at the other parameters characterizing users, it has been noticed that these users have also the highest number of clean characters and valid words.

A different view of the result of cluster analysis is given by the scatter plots of Figure 2, which show the characteristics of user profiles subdivided among groups. Each point represents

	cluster 1	cluster 2	cluster 3
	96,691 profiles	14,178 profiles	8,973 profiles
comments	75.59	23.1	15.12
links	1.57	3.79	1.86
words	19.52	28.22	100.28

TABLE III
CENTROIDS OF THE CLUSTERS.

the projection in the two-dimensional space of a user profile, expressed by two of the parameters used for its characterization. Profiles belonging to the first, second and third cluster, are represented in figure by plus signs, circles, and triangles, respectively. Note that, in order to make figure more readable, high values are not plotted. Looking at cluster one, it can be noticed that it contains users having a high number of comments in their profiles, but "poor" in terms of number of words and links. To the second cluster, belong users having a small number of comments, but a very high number of links. Finally, third cluster groups users having few comments, which however contain a high number of words. We can conclude that the minority of users have a small number of comments in their pages, but written in a formally correct way. Moreover, when very popular users attract many comments from friends, the content is "poor" in external references.

A. Analysis of external references

When writing a comment, a user can insert HTML tags in order to specify external references to Web pages, images and animated objects. As already seen, only about 9% of users have at least one link inserted in comments in their page. However, it is interesting and important to see which external sites are accessed and how accesses are distributed.

In the total number of comments considered, a total number of 48,317,140 references to images have been found, referring to 468,564 different sites (see Table I). Two sites, namely, *www.photobucket.com* and *www.imageshack.us* cover 50% of references. Moreover, the 18 most popular sites are responsible for 70% of the total number of references.

Comments contain also 47,876,519 links to external Web sites, and 260,625 different sites are accessed. Moreover, the first two most common sites, namely, *www.msplinks.com* and

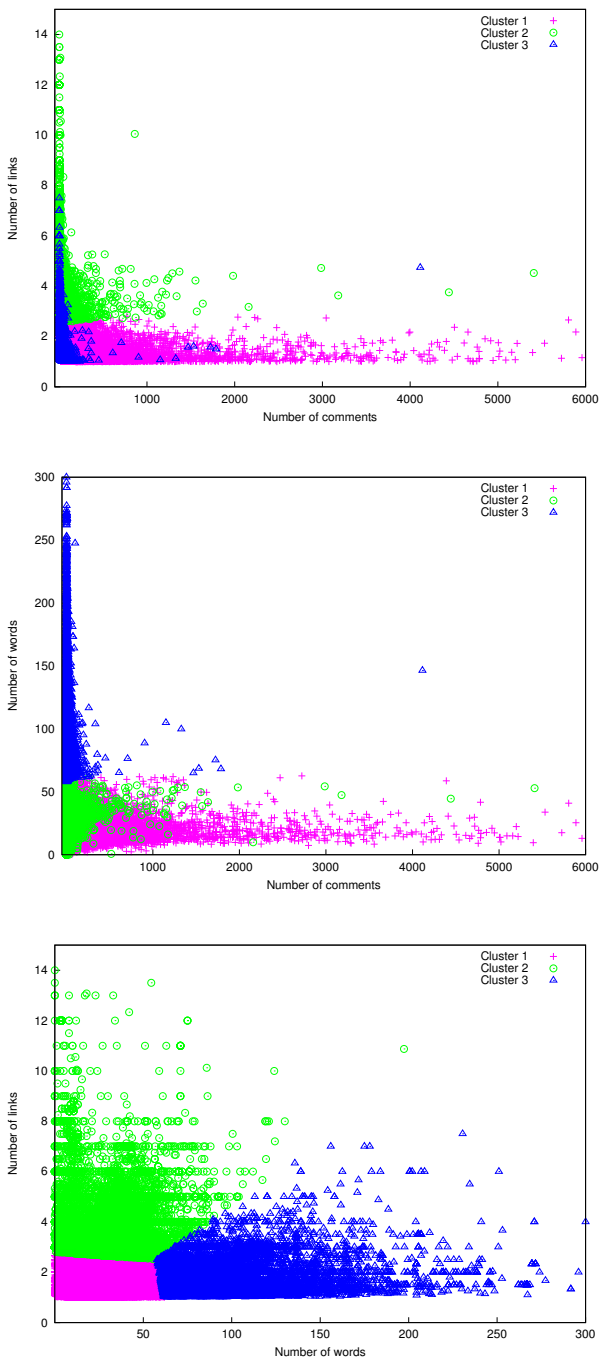

Fig. 2. Profile characteristics subdivided per group.

www.photobucket.com, account globally for 60% of references. Msplink is a link-redirect system started in mid 2007 to try to help stop spam and hackers on MySpace. Hence, accesses to *www.msplinks.com* hide accesses to external resources. Table IV shows the most popular sites used for referencing images and Web pages.

Animated, audio and video objects have also been identified in comments. A total amount of 2,771,047 objects are inserted by accessing 6830 sites. The most used site

Images	
www.photobucket.com	43.1%
www.imageshack.us	7.3%
lc.fdots.com	4.0%
myspacecdn.com	2.7%
www.glitter-graphics.com	2.2%
Web pages	
www.msplinks.com	53.2%
www.photobucket.com	6.5%
profile.myspace.com	5.3%
www.myspace.com	3.2%
www.imageshack.us	1.3%

TABLE IV
TOP 5 MOST POPULAR SITES ACCESSED IN COMMENTS.

is *www.youtube.com* accounting for 51% of the references, followed by *lads.myspace.com* with a percentage of 12%, by *static.slide.com* and by *mediaservices.myspace.com*. These four sites globally account for 73% of references to objects.

Another type of reference as also been discovered. Indeed, links and images, which are inserted as plain text, still remain after removing HTML tags. These references are simply read by users and cannot be accessed, but reflect a sort of communication which cannot be ignored. The number of this type of references is not negligible. 665,879 different links have been recognized in comments, giving a total amount of 13,642,564 occurrences.

B. Analysis of the words

On average, a comment contains 248.5 characters, but, after removing special characters, HTML tags, and punctuation, just 132 characters remain: this "overhead" of 53% is due so the inclusion of content other than words, such as, references to external document, and to special punctuation.

Each comment contains on average 26.1 words, and 90.2% of them belong to one of the selected languages, that is, are considered "valid". Within these, English words account for 97.18%, and are on average 23 per comment.
Words belonging to at least one dictionary have been classified as valid, whereas non-valid words could be due typo, as well as belong to the "Internet language", namely slang dictionary and emoticons. Table V summarizes overall characteristics of comments in terms of number of unique valid and non-valid words, total number of occurrences, and average per comment.

	unique	total	average
words	22,426,568	7,869,571,226	350.9
valid words	441,436	7,155,150,000	16,208.8
non-valid words	21,985,132	714,421,226	32.5

TABLE V
VALID AND NON-VALID WORDS IN COMMENTS.

Among valid words, "rare" and "common" words can be easily identified, according to their frequency. About 101,000 words, namely 23% of the words, can be classified as "rare", appearing just once. Moreover, "common" words, that is, very frequently used, are very few; 1699 very used words, are

only 0.38% of unique words, but cover 90% of occurrences. Figure 3 shows frequency of most used words; the most popular word appears about 300 million times.

Non-valid words are ”expressions” in an informal slang language, and are distributed in a clearly different way with respect to valid ones. Non-valid “rare” words, which appear just once, are many more than valid ”rare” words, accounting for 69.3% of non-valid words. Moreover, 89.7% words appear less than 7 times. The 166 most frequent words account for 30% of total occurrences, and the 1002 most frequent for 50%.

From these data, a model of user comes out using a poor, absolutely not variegated recognized language. Moreover, the number of unrecognized words is impressively high: in 41% of comments more than 10% of the words contained in the comment are non-valid.

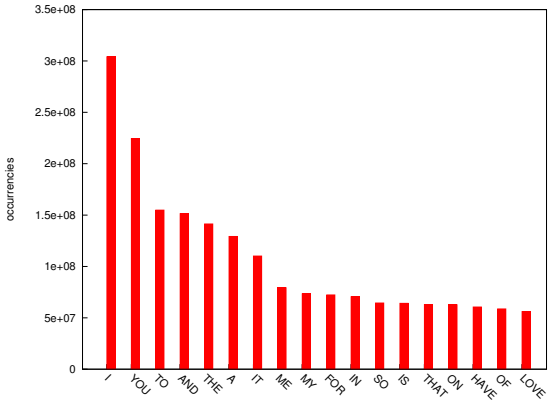

Fig. 3. Frequency distribution of most used words.

A few words have been recognized as belonging to more than one of the selected language. Table VI summarizes outcomes.

language	unique words	total
unique	378,637	2,962,720,000
English	100,404	2,800,680,000
French	42,413	30,627,358
Spanish	53,227	18,714,384
German	45,619	45,725,760
Italian	79,360	10,308,068
Portuguese	18,933	7,964,743
Dutch	38,681	48,699,522
2 languages	47,078	1,859,274,277
3 languages	11,384	797,205,923
4 dictionary	2,965	406,560,549

TABLE VI
COMPOSITION OF COMMENTS IN TERMS OF LANGUAGES USED.

English is clearly the most used language; 121,228 words have been classified as English, accounting for a total of about 7 billion words, that is, 96.99% of the total number of words. Moreover, it has also been noticed that the first 36 most popular words are English and account for about 2.65 billion of total occurrences, and that the number of English words in a profile is highly correlated with the number of words and of “clean” characters.

IV. CONCLUSIONS

This paper presents an analysis of MySpace comment structure. A huge amount of user profiles have been analyzed in order to extract and characterize the content posted in user comments. A comment has been described by means of parameters concerning its length, the language used, and the content inserted through external references.
Basic statistics are derived, and typical user behaviors have been identified, depending on the number of comments and on comments composition. This can be very helpful for social network analysis systems that apply either statistical or semantic analysis techniques to the comments they extract with keyword search. Future research could include comparison with other popular social network sites, such as, Facebook, and the analysis of profile evolution over time. Further analysis on the external links managed through msplinks is also required. Moreover, the analysis on the external sites accessed through comments could be extended to a deeper analysis of the different objects referenced, helping in discovering the most referenced multimedia content.

REFERENCES

[1] Myspace homepage. [Online]. Available: http://www.myspace.com
[2] Comscore web site. [Online]. Available: http://www.comscore.com
[3] Research-write. [Online]. Available: http://www.research-write.com/2010/02/social-networking-by-the-numbers.html
[4] R. Kumar, J. Novak, and A. Tomkins, “Structure and evolution of online social networks,” in *Proc. of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006, pp. 611–617.
[5] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” in *ACM Internet Measurement Conference (IMC’07)*, 2007, pp. 29–42.
[6] B. Saha and L. Getoor, “Group proximity measure for recommending groups in online social networks,” in *Proc. of the Second SNA-KDD Workshop on Social Network Mining and Analysis*. ACM Press, 2008.
[7] W. Willinger, R. Rejaie, M. Torkjazi, and M. Valafar, “Research on online social networks: Time to face the real challenges,” *SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 3, pp. 49–54, 2009.
[8] C. Aguiton and D. Cardon. (2007) The strength of weak cooperation: An attempt to understand the meaning of web 2.0. Available at http://ssrn.com/paper=1009070.
[9] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, “I Tube, You Tube, Everybody Tubes: Analyzing the World’s Largest User Generated Content Video System,” in *ACM Internet Measurement Conference (IMC’07)*, 2007, pp. 1–14.
[10] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, “YouTube Traffic Characterization: A View From the Edge,” in *ACM Internet Measurement Conference (IMC’07)*, 2007, pp. 15–28.
[11] M. Halvey and M. Keane, “Analysis of online video search and sharing,” in *Proc. of the 18th Conference on Hypertext and Hypermedia (HT07)*, 2007, pp. 217–226.
[12] X. Cheng, C. Dale, and J. Liu, “Characteristics and Potential of YouTube: a Measurement Study,” in *Peer-to-Peer Video - the Economics, Policy and Culture of Today’s New Mass Medium*, E. Noam and L. Pupillo, Eds. Springer, 2008, pp. 205–217.
[13] G. Urdaneta, G. Pierre, and M. van Steen, “Wikipedia Workload Analysis for Decentralized Hosting,” *Computer Networks*, vol. 53, no. 11, pp. 1830–1845, 2009.
[14] E. Cohen and B. Krishnamurthy, “A short walk in the Blogistan,” *Computer Networks*, vol. 50, no. 6, pp. 615–630, 2006.
[15] J.Caverleee and S. Webb, “A large-scale study of myspace: Observations and implications for online social networks,” in *Proc. Int. Conf. on Weblogs and Social Media (IICWSM 2008)*, 2008.
[16] M. Thelwall, “Myspace comments,” *Online Information Review*, vol. 33, no. 1, pp. 58–76, 2009.

[17] M. Thelwall and D. Wilkinson, "Public dialogs in social network sites: What is their purpose?" *Journal of the American Society for Information Science and Technology*, vol. 61, no. 2, pp. 392–404, 2010.
[18] L. Massari, "Analysis of myspace user profiles," *Information Systems Frontiers*, no. 10.1007s10796-009-9206-8, 2010. [Online]. Available: http://www.springerlink.com/content/t4361h5n77037247/
[19] J. Hartigan, *Clustering Algorithms*. John Wiley & Sons, 1975.
[20] The R project. [Online]. Available: http://www.r-project.org/