Consolidating Data For Forming An Information Image Of Geospatial Objects From Heterogeneous Social Media

Iryna Khmil
Dept. of Social Communication
and Information Activity,
Lviv Polytechnic National University
Lviv, Ukraine
Khmil.iryna@gmail.com

Abstract— In this paper the textual content and other attributes of heterogeneous social media are analyzed for extraction of consolidated data for forming an information image of geospatial objects.

Keywords—social media; LBSN; content; geospatial object;

I. Introduction

Within the study and analysis of virtual communities separately released a service segment which contains geographical data about real-world geographical objects, and accumulates the result of the activity of the people shown in these communities - Location-Based Social Networks.

Most studies about users and their geographical location are objects of study. The main areas are:

- Analysis of social networks with the making of voluntary information [1];

- Analysis of large volumes of text and evaluative information in order to obtain specific recommendations about visiting geospatial objects;[7,8,9];

- The study of the behavior and trajectories of user and comparison it with others. Consequently algorithms of comparison, clustering and determining the similarity between users are created. These studies allow bringing a pair or set of users that can be useful to each other or determination one user as local expert on specific geospatial object [6];

- Determining correlations between social and geographical connections of users.[10]

However, all studies are conducted on the basis of data obtained from the array of one LBSN or more similar in their information content and must contain the precise geographical data geospatial object. Almost reflected social network segment that does not have the exact location as the primary attribute (longitude, latitude; postal address), but it is possible to determine relation between text and geospatial object.

II. The choice of social networks

The social environment of the Internet is based on any apparent activity, social activity of a significant number of Internet users, particularly in different social networks, social services narrower purpose. Such an environment is the source of heterogeneous information, collection, selection, processing and analysis of which allows to create a new layer for the GIS and other resources that require geographically marked information.

While such information can not be immediately called geographic because it does not contain specific geographic coordinates, but has a geographic track [3,4]:

- Intentional geotaging of their location with commentary in various kinds of social networks;

- Connect to messages` texts in microblogs automatical determining the location of the author;

- LBSN, purpose of which is to provide guidance on the crowded places in real life.

Using these services makes possible to get information in real time about the social hot spots in the real environment. Such information is aging fast, but in real time it is one of the most complete and accurate according to the information request Internet users and their situational awareness.

Questions of authenticity of information geotaging are decided by using the search engine placement, which are equipped with mobile devices and with access to the Internet (not all devices have navigation systems are configured directly via satellite).

Unlike the approaches to creating volunteer geographic information [1], there is focus on motives of generating content in the above listed social services are not creating it geographic information and create information in order to promote their social profile (social networking, micro blogging), the exchange of useful information etc.

Also volunteer services of geographical information do not allow identifying the time stumps of creating content that is required to analyze and monitor changes of image information objects.

Geographical track of the author should also be determined from information text content of the social environment of the Internet, where we cannot get reliable data for the location of search author inception post any kind of Internet activity.

It just raises the question whether the author of location information is crucial in the analysis of social activity in the real environment in relation to certain geographical individually selected object, or a reference to an object in the text will be more informative and higher compliance with information requests.

That is why for further analysis of the information image of geographic object was chosen three types of virtual communities: forums general general social communities (Facebook, VK, Twitter), Location Based Social Networks. Although they differ in the way of detecting the geographic object and the relevant posts, but the method was bringing data from such heterogeneous environments in a unified view.

III. Ways to obtain and unify data from hreterogenous social media.

Despite the heterogeneity of social media information described below can be obtained from all types of communities.

1. *Search for relevant posts.*

To find the relevant to the geographical object posts in forums and communities of this type need to compile a dictionary of all possible names of the object. In particular you need to consider how it will look at the name transliterated into Latin alphabet (with the possibility that the original name is displayed hieroglyphs, Arabic script, etc.). Knowledge of the location may help when the same or similar in sound / spelling determine GsO located on different continents (US state Georgia and country Georgia in English are the same in writing); when the same name with different types of GsO (temple Münster and Munster town in Germany); when are the same names of settlements, even within one state. Similarly with institutions whose names can be duplicated in different cities. Questions transliteration of proper names and place names in the search for information is urgent. Each language of non-Latin alphabet has its own rules of transliteration into Latin. Also almost impossible to have knowledge of all variants of proper names in languages with Latin alphabet, for example: English. - Lviv, Polish - Lwow. Accuracy determined by finding the appropriate information on the GsO when the design contains a request two or more words. For example, when you enter the name New York we get the city of New York, not the GsO that contain the words «New» and «York» - New Orleans, New Zealand, York.

When we search on forums or in social networks problem arises in part of working with relevant textes.

In LBSN such problems should not occur.

2. *Identification of the social profile of the author.*

The most important characteristic of the author is to identify his name on the Internet (nickname or real - tangible impact on research it has).

In networks such as forum such identification possible on following of other posts within discussion or other discussions.

In general and specialized social networks social profile is closely linked with the next point - when the profile can be most accurately read to identify of the author's social profile and its activity in respect of the given geospatial object [5].

3. *Identification of implication (residence and consumence)*

Belonging to group in forums defined (consumption or residence) by markers - words, phrases, language structures that uniquely identify the characteristics of the author. Characteristics of author is unique only in that time in which the published notice and may vary over time. This change is evident in the analysis and geographic information track based on social activity in a social network. So we get retrospective data of one author who organizes the temporal database.

In this study, we are not interested in changing of GsO information image of author as a real entity, but as subject to certain social network defined characteristics, because each message gets some metadata regarding authorship.

4. *Identification of time stamps (difference between time of visiting and time of publication of experience).*

Important in identifying the posts adequacy is recording time stamps of posts creation and identification of physical time stay or receive any other experience by the author. This is because the characteristics of some objects are changed more frequently than others, so the information may be outdated character within the analysis of the current image.

5. *Analysis of the type of expressing experience in posts.*

There is considered 4 types of expressing the experience of consumers within the same post:

- Descriptive - express knowledge of geospatial object in narrative form. There is expressed the possibility of obtaining certain services, the possibility of action or lack of them.

- Declarative - expressing the most significant and important thoughts. There is text accompanied by an additional emotional stress, use of exclamation marks or emoticons. Most often occurs at the beginning or end of the post, can be in the same sentence with the recommendation;

- Recommendation - expression contains a recommendation or a call to action (or inaction);

- Question - contains consumers' questions regarding geospatial objects in order to get direct experience. In this type of statements interest and the demand for certain types of information is expressed.

6. Posts separation into components according to analysis:

- The concept of first level, second level concepts$

- Help word;

- Markers of consumence and residence;

- Markers of accuracy, adequacy;

- Descriptors;

7. Giving coefficients to each descriptor.

It is used to build a hierarchy that reflects the image of geospatial information object.

Building a hierarchy of representativeness values carried out by each of the descriptors in all.

Top of the hierarchy is a concept - a term that denoting the text geospatial object. Because it is a mandatory attribute relevant post the value of a concept is always 1. The first level of the hierarchy will amount descriptors, which has a value of representativeness deviation of more than 0.01 of this descriptor with the highest value. The following levels are based on connections R, which have the following meanings:

R_h - couple descriptors found in a single sentence in one type of expression of one message;

R_{m1} - pair descriptor is found within one type of expression, but in different sentences;

R_{m2} - pair descriptor is found within the same sentence, but in different types of expression

R_l - descriptors occur within a message, but in different sentences for various types of utterances.

If the deviation values of representativeness of the analyzed descriptor from the descriptor one level in the hierarchy is less than 0.01, but relation with either one of them is Rl, but with a higher level descriptor (or concept) is another type of relationship, the handle is put in this new level node.

REFERENCES

[1] Goodchild M. F. Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. / M.F. Goodchild // International Journal of Spatial Data Infrastructures Research, 2(1), 2007,pp. 24–32.

[2] A. Stefanidis, A. Crooks, Ja. Radzikowski Harvesting ambient geospatial information from social media feeds, GeoJournal, Volume 78, Issue 2, April 2013, pp 319-338

[3] I. Khmil Content of social media as a sourse of geographic infommation, Conference Proceedings 'MKCI' Kharkiv, Ukraine, vol. 9, April 2015, pp. 54-55.

[4] I. Khmil Analysis of the interaction of geospatial objects` image information components in virtual communities. Management of Development of Complex Systems, Issue 21, 2015, pp. 87-91.

[5] A .Peleshchyshyn, Yu. Serov, S. Fedushko Development of registration and validation algorithm of personal information of Web community members, Computer Science and Information Technologies, Vol. 686, 2010, pp. 238-244.

[6] Yu Zheng Location-Based Social Networks: Users, Computing with Spatial Trajectories , 2011, pp,243-276.

[7] Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.Y.: Recommending friends and locations based on individual location history. ACM Trans, 2011, Web 5, 5:1-5:44.

[8] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM 2011.

[9] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative Location and Activity Recommendations with GPS History Data. In Proceedings of the 19th International Conference on the World Wide Web, WWW 2010, pages 1029–1038.

[10] Eagle, N., de Montjoye, Y.A., Bettencourt, L.M.A.: Community computing: Comparisons betweenrural and urban societies using mobile phone data. In: Proceedings of the 2009 InternationalConference on Computational Science and Engineering - Volume 04, 2009, pp. 144–150.