CO-461

# MINING USER GENERATED WEB CONTENT TO CHARACTERISE GEOGRAPHIC FEATURES BEYOND TOPOGRAPHICAL ASPECTS

*MASSIOT L.(1), ABADIE N.(2), BUCHER B.(2)*

*(1) Centre National des Arts et Métiers, PARIS, FRANCE ; (2) Université Paris Est, IGN-COGIT, SAINT-MANDÉ, FRANCE*

## INTRODUCTION: GEOGRAPHIC INFORMATION AND USER GENERATED WEB CONTENT

Geographic databases have long been produced exclusively by experts. These were either experts from a dedicated thematic field or cartographers. Lately, new kinds of digital geographic information have emerged based on collaborative web sites or applications that merge user contributions, like Wikipedia, Geonames or OpenStreetMap (OSM). In these contents, some geographic information is explicitly represented and some can be mined.

The growth of such user generated content (UGC) is most relevant to geographic information management because the production processes in National Mapping Agencies (NMAs) and in UGC have different advantages and constraints. NMAs have made choices to represent a territory on map series and more lately in reference geographic databases. Their models try to answer to most needs requiring maps or reference data, i.e. representation of real world features with their characteristics to position phenomena in geographical space or evaluate spatial relationships between features. A major choice of NMAs to meet such challenges is to organise the representation of the geographical space into different levels of details. They also commit to respect a specific update frequency and quality criteria. As a counterpart, their models and updates are not very flexible. On the contrary, UGC can focus on specific applications, like in UCrime. Besides, contributions to models or instances can be made at anytime, usually with a moderation step. Hence their representation is globally very flexible, like in Wikipedia. A counterpart is that consistency and homogeneity are challenging aspects for UGC.

To our opinion, these differences in NMA and UGC production context make these information sources somewhat complementary. UGC has a more flexible creation process and can be updated more rapidly than databases produced by public agencies. This is why UGC content is often seen as an interesting way to generate alerts for NMAs to update their own contents. Besides, the fact that any one can theoretically contribute to most UGC make them an interesting tool to analyse what kinds of knowledge are relevant to describe a territory. For example, even though OSM shares some descriptive elements of the geographical space with NMA databases, like roads or street names, it also contains other feature types like cash machines. Several works have explored the possibilities of UGC to acquire new features or to enrich features description typically with footprint information (Dahinden and Sester 2009) (Dahinden 2009). Mining tags in social applications is another opportunity to do that (Becker et al. 2009) (Rattenbury and Naaman 2009).

The work presented in this paper does not aim at extracting instances but rather at extracting model-level information (classes and attributes). To prototype and illustrate our approach, we have selected a specific kind of geographic objects in Wikipedia and in the French NMA large scale database IGN BDTopo®: mountain huts. The following of the paper describe our work to select and mine what can be called a user generated model for mountain huts description and to compare it with a NMA feature type and attribute types for mountain huts. 'Select and mine' means that some items of this User Generated Model (UGM) are explicit in Wikipedia and that others have to be extracted. We firstly focus on a first level UGM: identifying classes of mountain huts in Wikipedia and compare them with BDTopo® data. This is described in the following section. Then we refine this first level UGM by examining what distinguish items one from another beyond the extracted classification scheme. This is described in the section after. Eventually, we draw some perspectives for further works.

## FIRST LEVEL USER GENERATED MODELLING ELEMENTS FOR MOUNTAIN HUTS IN FRENCH WIKIPEDIA

In Wikipedia, a mountain hut is represented through a dedicated page: a mountain hut article. A Wikipedia contributor who writes an article on a specific mountain hut is encouraged to use a dedicated model: a set of fields adapted to the description of a mountain hut and called the 'mountain hut Wikipedia infobox'. The Wikipedia explicit model for mountain hut is made up of the following attributes: altitude, massif,

country, region, department, manager, opening days, number of beds, coordinates (point). The NMA IGN BDTopo® model for mountain hut is the following: coordinates (point), name, lineage information.


*Figure: The rectangle on the right presents values for the infobox attributes for this specific mountain hut. Highlighted words corresponds to internal links within Wikipedia that relates words used in the hut description to Wikipedia articles defining these words.*

Besides documenting this infobox, the contributor is also expected to freely describe his mountain hut. We are interested in analysing these "free" descriptions to find if people use specific kinds of information to characterise mountain huts and distinguish them one from the other. To do so, we do not use every word in the description but only internal links. Internal links are assumed to relate to concepts important enough to have an article in Wikipedia and somehow connected to the current mountain hut. The space where we calculate similarity between two mountain hut articles was firstly made up of every internal links found in French mountain hut articles. The main encountered difficulty is related to managing the dimension of the descriptive space with respect to the number of items. Indeed there was 77 items for a space of 377 dimensions. In such a space, all items are globally different. We have reduced this space by keeping internal links that were also categories, i.e. internal links that were references to Wikipedia thematic index called Wikipedia category graph. We have also removed internal links which were not meaningful for the clustering, such as those which appeared in all articles and those which appeared in only one article. In the resulting space, for each mountain hut article, the value of its coordinate on an "internal link dimension" is the number of time this internal link appears in the article. We use the Tanimoto coefficient to compute similarity between these vectors (see formula 1). The similarity between two articles is the similarity between their vector representations in this space, weighted by the TF-IDF (Term Frequency – Inverse Document Frequency) of the internal links in the set of mountain huts articles.

$$T(A,B) = 1 - (A.B)/(||A||2 + ||B||2 - A.B)$$

Practically, Wikipedia mountain huts instances are automatically extracted from Wikipedia dump. The native database management system eXist has been used to organise and manage a Wikpiedia XML database. It comes with a free text search system, Lucene, which has been used to index articles based on their titles and infobox. The coordinates provided by infobox properties have been used to select mountain huts that were located in France. The clustering is based on Lingpipe system and the "within cluster scatter" (wcs) measure. The number of clusters z is chosen so that the evolution of the wcs measure between z clusters and (z-1) clusters is the greatest. We have obtained seven clusters.

Resulting clusters are then characterized. A map of these clusters (see fig. 2) shows that mountain huts belonging to a cluster have spatial relationships: they belong to the same mountain chain or they are close to the same hiking track. For each cluster, we have analysed internal links that appear in 60% articles of the cluster. Generally, these links were all geographical categories but for the categories “sport en haute savoie” and “hiking”. The exception was made up of a singleton corresponding to an observatory “Refuge Vallot”.

**SECOND LEVEL USER GENERATED MODELLING ELEMENTS FOR MOUNTAIN HUTS IN FRENCH WIKIPEDIA**

The section after examines internal links that have not been used in the clustering step. We aim at identifying semantic categories of such internal links, i.e. natures of descriptive items that distinguish one mountain hut from another. From the initial set of internal links we extract a subset composed of each link that appears in one mountain hut and not in the other mountain huts belonging to the same cluster. We call such internal links "typical links". The articles corresponding to these typical links have categories. Instead of examining the links themselves as in the preceding step, we examine these categories. And instead of clustering articles as in the preceding step, we cluster these categories. In other words, our objective in this step is to extract natures of descriptive items that are relevant to distinguish one mountain hut from another.

Practically, for this analyse, we have used the Wikipedia categories graph like a ontology. The clustering of typical links categories is still based on Lingpipe system. It has been performed based on a semantic distance measure, computed as the additive inverse value of the (Wu and Palmer, 1994) semantic similarity measure. We have chosen this semantic similarity measure since we wanted to obtain higher similarity scores for pairs of categories having subsumption relationships than for those having sibling relationships.

Interestingly, at this level, non spatial categories appear. Among the created clusters we can notice references to companies such as "Météo France", the French power supply company ("Electricité de France"), the French railways company ("Société Nationale des Chemins de Fer") or the French National Centre for Scientific Research ("CNRS"). More outstandingly, famous people are also often cited to characterise mountain huts, either because they have contributed to the construction or the inauguration of that hut or because they have a significant relationship with it. More specifically, the sub categories of famous men that are used to characterise French mountain hut are: politician, minister, President of France, International Olympic Committee member, members of the Académie Française, geographer, photographer, natural scientist, artist, engineer, astronomical scientist, meteorologist, mountaineer, Pyrenean mountain climber, alpine guide, Belgian monarchs and Knights of the Garter.

**DISCUSSION**

Wikipedia can be seen as a most interesting information repository to analyse what kinds of knowledge are relevant to describe a territory. The work presented in this paper experienced the use of Wikipedia to extract a user generated model of a specific type of geographic object, the mountain hut, beyond Wikipedia categories and infoboxes.

It is a promising technique and there remain a lot of issues to solve. In our study, extracting pieces of information related to mountain hut features was an easy task because there is one article for each of them and these articles can easily be accessed thanks of titles usually containing the word “hut” (“refuge” in French) and most of them have the corresponding infobox. The second similarity measure used in our study is based on the Wikipedia category graph considered as a ontology. Yet, this is not a ontology, but rather a folksonomy. Future works concentrate on refining the distance used to cluster descriptive items

and to cluster studied features. For the first, we aim at using other web ontologies, aligned with the Wikipedia category graph. For the second, (Andrienko and Andrienko 2010) measure could be tested.

One perspective of this work is to assist users who want to build a structured database to design the model for this database so as to ensure its consistency with existing content, NMA and Wikipedia (Brando et al. 2011).

**AKNOWLEDGEMENTS**

The authors wish to thank gratefully Professor Fouad Badran from the Conservatoire National des Arts et Métiers for his assistance in the selection of a distance to cluster internal links in Wikipedia French mountain hut articles.

**REFERENCES**

Brando, C. Bucher, B., Abadie, N., 2011, Specifications for User Generated Spatial Content, in proceedings of the AGILE conference, Utrecht

Dahinden, T., Sester, M., 2009, Categorization of linear objects for map generalization using geocoded articles of a knowledge repository, in proceedings of the international workshop Presenting Spatial Information: Granularity, Relevance, and Integration, held in conjunction with the Conference on Spatial Information Theory, COSIT, France

Dahinden, T., 2009, Localization of uncertain and fuzzybordered areas by geocoded articles of a knowledge repository, in proceedings of the 24th International Cartographic Conference, Chile

Andrienko, G., Andrienko, N., 2010, Sammon's projection for clustering complex geographical objects, in Purves, R. (Ed.) et al.: GIScience 2010: Sixth International Conference on Geographic Information Science, Zurich

Rattenbury, T., Naaman, M., 2009, Methods for extracting place semantics from Flickr tags. ACM Trans. Web, vol. 3(1), Article 1

Becker, H., Naaman, M., Gravano, L.. 2009, Event Identification in Social Media. in Proceedings, 12th International Workshop on the Web and Databases: colocated with ACM SIGMOD

Wu Z. & Palmer M. (1994). Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics. p. 133-138.