Will You Like this Place? A Tag-Based Place Representation Approach

G.B. Colombo, M.J. Chorley, V. Tanasescu, S.M. Allen, C.B. Jones, R.M. Whitaker
School of Computer Science & Informatics, Cardiff University, UK
{*g.colombo, m.j.chorley, v.tanasescu, stuart.m.allen, c.b.jones, r.m.whitaker*}*@cs.cardiff.ac.uk*

***Abstract*—Increasingly popular location-based services can monitor our geographical positions in real time and so can provide a fundamental source for capturing users feelings and personal attitudes towards a particular place at a particular time. We propose a novel procedure for the representation of places through weighted tag-lists based on user reviews on these type of services. In our method the resulting lists can be built according to different criteria aiming to highlight differences and similarities among locations that can be in geographical proximity, belong to a similar type/category, or be included in the personal mobility history of a specific user.**

***Keywords*-human mobility; keyword extraction; sense of place; location-based services;**

I. Introduction

Location based social-networking services have become popular among users of portable internet-enabled devices such as smartphones and tablets. Through these services consumers can leave tips or reviews for other users expressing personal opinions and 'feelings' about a place in an informal way, free of conditioning. Potentially, these user provided opinions and feelings may not be consistent with each other, and may not focus on the same aspects of a place since they will reflect the 'perception' that individual users have towards a particular location, rather than a more objective description of it.

Within pervasive and mobile computing there is a desire to utilise the wealth of information made available by (and accessible to) mobile devices in order to create autonomous systems capable of providing useful content and place recommendations to users. The underlying idea of this research is that by aggregating the information included in personalised reviews of locations we can obtain a representation of a place that depicts it as a projection of (different) users perceptions in line with the notion of *sense of place*, which emphasises the characteristics that make a geographical location special or unique for a particular individual [1]. The unique approach that we adopt is to build a tag-based representation of a place by extracting keywords from a collection of online documents. Personalisation of content provision based on the situation in which the user finds themselves is well suited to a tag-based approach, allowing quick access to a wealth of situated location-specific content.

II. Related Work

With the advent of mobile devices and location-based services that can monitor our geographical positions in real time, mobile and online services related to places are even more popular, evolving from an initial adaptation of online maps and navigators towards services more oriented to provide reviews and personalised recommendations such as *Yelp* and *Qype*[2], to others that combine location and user mobility with a social networking component, for example *Foursquare*, *Flickr*[3], and *Google+ Local*[4]. All of these services have evolved towards a place representation that is more related to the individual needs of users, with users being at a particular location at a particular time often making use of tags, annotations and other user generated content [2].

[3] analyses how Foursquare users exploit tips, dones and to-dos in relation to different behavior

[2]http://www.yelp.co.uk/, http://www.qype.co.uk/
[3]https://foursquare.com/; http://www.flickr.com/
[4]http://www.google.com/+/learnmore/local/

978-1-4244-9529-0/13/$31.00 ©2013 IEEE

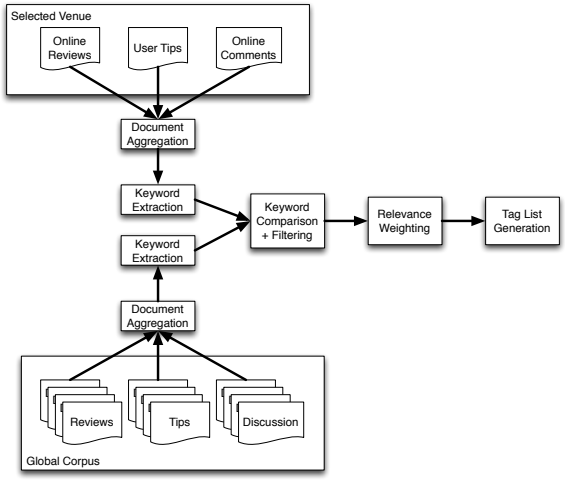


Figure 1. Tag-list generation process

profiles. Similarly, [4] applies a clustering algorithm to the *Foursquare* database of two cities in order to detect communities of users visiting similar categories and compare them to the distribution of urban areas and neighborhoods. This can lead to a re-design of the concept of an urban area as a characterisation not only based on the the types of places and locations in its neighbourhoods, but also on the people who actually live within the area and make it part of their daily routine with their own (possibly contradictory) perceptions of it [5].

The research presented in this article aims to provide further contributions in this direction. Specifically we present an initial proof of concept implementation of a tag-based place representation which can be used to compare, contrast and recommend places and content to users[5].

III. Methodology

Online and mobile location-based services often associate the representation of a specific venue with one or a limited number of keywords (e.g Google+ Local, Yelp). However, these terms primarily represent general characteristics that are also shared with other venues of the same category (even when they go beyond the category itself). We seek to propose a method for a keyword representation of places and venues that will more realistically reflect the needs and the feelings of people as single individuals.

Figure 1 shows the tag-list generation process. It consists of several steps during which documents from different sources are aggregated, reduced, compared and finally weighted to provide the final weighted tag-list:

1) **Document Aggregation** The set of documents describing a venue within an area is aggregated.
2) **Corpus Construction** A set of documents to compare the venue against is constructed. This may be comprised of many different types of document, depending on the need of the system including:
   a ***Global corpus*** documents describing all venues within a given area are aggregated;
   a ***Category corpus*** aggregating documents describing all venues with a specific type;
   a ***Personalised corpus*** aggregates only the documents describing venues related to a user's personal history. This can be based, for example, on user mobility patterns or on some aspect of their personality.
3) **Keyword Extraction** The corpus of documents is filtered for common stop-words and words with low frequency.
4) **Keyword Filtering** The keywords from the venue corpus and the comparison corpus are compared for similarity to identify keywords specific to the venue under comparison.
5) **Tag-List Generation** Keywords are weighted according to their frequency of appearance within the corpuses.

Sources currently used are online reviews and tips from Foursquare, Google+ Local, Yelp, Qype, and text extracted from the venue websites. Tags given for a venue are often related to its basic features, such as its type and category (e.g pub, coffee shop, restaurant), but they can also reveal other distinctive characteristics. For example, some coffee shops may be preferred because they serve particular kinds of dietary food, some others for services provided (e.g Wi-Fi), and some venues

[5]This research was funded by RECOGNITION an EC - FP7 Future Emerging Technologies project

may be liked because of their particular atmosphere and the social environment surrounding them. The corpus used for comparison obviously has an effect on the resulting tag list, and it is here that much user personalisation can be carried out.

Term Frequency (*TF*) is a simple weighting scheme for keywords in a document, that uses the bag of words model. *TF* assumes the weight of a keyword to be equal to the number of occurrences of term t in document. Term frequency alone however has little discriminating power in a themed corpus, as some keywords will probably be found in all documents. For example the word 'beer' will be found in many or all pub reviews. The idea is therefore to adjust term frequency using the count of occurrences of the term in the whole collection. This measure is the document frequency *DF*, or number of documents that contain a term t. To use *DF* to scale the term frequency, the inverse document frequency *IDF* of a term t is defined as

$$idf(t) = \log(N/df(t)) \quad (1)$$

where N is the number of documents in the corpus. The *TF-IDF* weighting scheme assigns to term t a weight in document d given by:

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \quad (2)$$

The weighting has the following characteristics [6]:

- highest when the term occurs many times within a small number of documents
- lower when the terms occurs fewer times in a document, or occurs in many documents
- lowest when the term occurs in virtually all documents

Once the ranked list is provided for each venue a similarity measure between two places can be computed using a standard *cosine similarity* in which the weighting of the individual keywords is provided by *TF-IDF*. This is a popular measure of similarity for text clustering, which captures a scale invariant understanding of similarity, so compensating for the effect of document length in a corpus [6], [7]. Cosine similarity is expressed as:

$$cs = (V(d1) \cdot V(d2))/|V(d1)||V(d2)| \quad (3)$$

In the extraction phase the text extracted from on-line review is tokenized, uncapitalised, stripped of punctuation, stop-words, non-English words and special characters, and part of speech (*POS*) tagged to filter verbs and adverbs. The Natural Language Toolkit *NLTK* provides easy tokenization, access to the Wordnet thesaurus for the elimination of non English words, as well as *POS* tagging[6].

IV. Results

As a proof of concept we have applied the weighted tag-list procedure to a number of venues in Cardiff, UK, focusing on two categories: 'pubs' and 'coffee-shops' (as classified by Foursquare). The venues were retrieved by taking two central points in selected Cardiff locations: the coffee shop 'A Shot in the Dark' and the pub 'The Pen & Wig' (among the most popular places for their type) and then considering all other venues within a circumference of a fixed radius (500m) around this centre. We will refer to these areas as $A1$ and $A2$ respectively with the latter representing a more central area and with the former more oriented to students and younger demographics.

A. Inner and outer similarity between categories

Our first experiment includes all the available venues in the definition of the corpus (Global corpus) and aims to detect differences in the relative similarity by considering either venues of the same or different type. We have retrieved a total of 50 venues of several different categories including 'coffee shops, pubs, breweries, bars, restaurants, clubs, pizza places, fish and chip shops, hotels, movie theatres, student centres and academic buildings'. We then calculate the weighted tag-list for each of the venues and compute *cosine similarity* between each pair (for ease of representation self-similarity has been set to null). We can generally expect higher similarity among venues of the same category. The diagram in Figure 2 visualises similarity between pairs of venues ordered by type/category. Although we can recognise peaks of high similarity among the same category, we can also note occasional peaks among

[6]http://nltk.org/, http://http://wordnet.princeton.edu

venues of different type, which is in contrast to the current recommendation criterion applied by many location based services that suggest similar places essentially within the same type-category only.

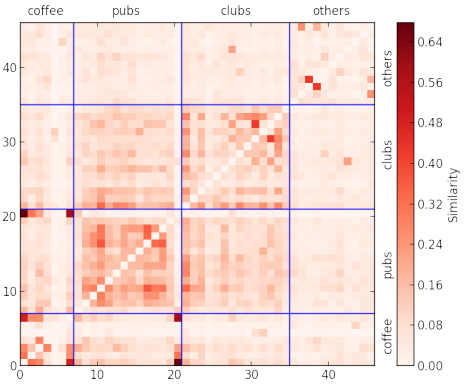
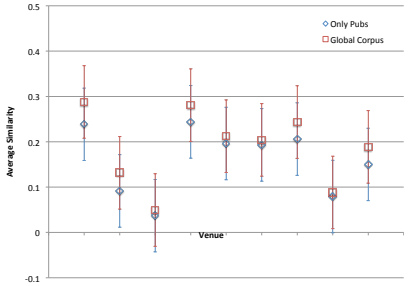


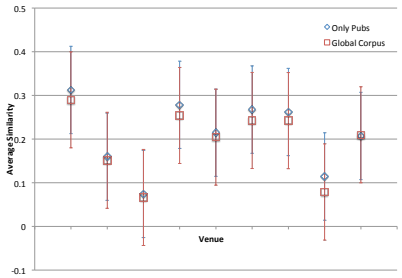
Figure 2. Pairwise Similarity of all Venues - Cosine similarity is higher within categories, but peaks of high similarity exist between different categories

B. Effect of different Parts Of Speech on different corpus choices

We have then considered the set of venues in the area $A1$ (the neighbourhood of 'The Pen & Wig' within a radius of 500m) and its subset of pubs $P1$ (the Foursquare API returns 9 pubs having online reviews this neighbourhood), Then we have computed the average cosine similarity that each of the pubs in $P1$ has with any other venue in the set $A1$ using either the Global or the Category corpus that only considers venues categorised as 'pubs' (see Figure 3(a)). The idea is that with the use of a corpus filtered by type the *TF-IDF* algorithm highlights (by weighting more) the keywords representing differences between places of the same category, thus penalising the terms representing the most *commonly shared characteristics* and resulting in a lower average similarity (from a One-Way ANOVA results are statistically significant at a p value 0.1). Figure 3(b) shows the same computation but in the case of using only adjectives as accepted lexical terms. Here the trend is reversed, with the differences between the averages no longer significant at the same



(a) all terms



(b) adjectives only

Figure 3. Similarity of Pubs - comparison between global and filtered corpus (by type) - all terms

significance interval (ANOVA analysis returns a p value $0.32 > 0.1$). A possible explanation is that when considering adjectives only we focus at terms that describing more the 'personality' and 'general atmosphere' of a venue. Since this can sometimes be more similar among places belonging to different categories rather than within the same one, we now reduce the 'negative' effect of these *commonly shared terms* when we consider venues belonging to the same category only (as we do with the Category corpus).

C. Self-similarity with different corpuses

We now consider again the subset of 'pubs' $P1$ in the neighbourhood $A1$ of 'The Pen & Wig' and the subset of 'coffee-shops' $C2$ in the neighbourhood $A2$ of 'A Shot in the Dark'. Table I shows the cosine similarity calculated for each of these venue between the weighted keyword list produced by including all venues (Global corpus) and those obtained by filtering tags either by category (column 1) or by considering a corpus based

on an individual user's mobility characteristics (column 2).

Venue name	Global vs. Category	Global vs. Personal
Coffee Shops - C2		
A Shot in the Dark	**0.8615**	0.7534
AJ's	0.9238	0.8725
Caffe Nero	*0.9147*	**0.1445**
Espresso Bar	0.9542	0.9432
Pubs - P1		
The Pen & Wig	0.9490	0.9131
Tair Pluen	0.9808	0.9573
Tynant Inn	0.9864	0.9649
Rummer Tavern	0.9549	0.8848
O'Neills	0.9745	0.9110
Varsity	**0.3425**	**0.2193**
Owain Glyndwr	0.9545	0.9326
The Central Bar	*0.6315*	*0.9355*
Duke of Wellington	0.9824	0.9426

Table I
COSINE SIMILARITY BETWEEN WEIGHTED TAG LISTS BUILT WITH GLOBAL AND FILTERED CORPUS

Global corpus vs. Category corpus. As expected, the results in column 1 show that all pairs of tag lists related to a specific venue are very similar (with similarity value greater than 90% in most cases). The main exception is one venue 'Varsity' (0.34) that is fact is not only a 'pub' but also has 'bar' and is a popular place for food, drinks, and nightlife in general. The use of a filtered corpus has the effect of penalising the terms that are shared within the pub category (in this case the words 'food' and 'pub') and vice-versa promoting those words that are not so typical of pubs (in this particular case the word 'bar' brought to top of the weighted list and having the highest weight when considering a filtered corpus, but being low in the list when all categories are included in the corpus). This results in a strikingly low similarity between the filtered and non-filtered keyword list. For the category of coffee shops the only venue that is below the 0.9 threshold of similarity is 'A Shot in the Dark' (0.86), which also does not completely comply with the traditional idea of a british coffee-shops (it has live music, it serves alcohol etc.).

Global corpus vs. Personalised corpus. We here use the actual preferences and behavioural characteristics of a specific user based on the history of the places visited to define the relevance and weighting of the weighted tag-list (i.e. if a user prefers to visit venues of a certain type the weighting of the keywords will highlight differences within places of that type). We show in the second column of Table I an example based on the Foursquare history of check-ins of one of the paper authors. We see that in general self-similarity between tag lists produced with the Global corpus and the Personalised corpus (col. 2) appears lower then the one obtained using Global and the Category corpus (col. 1). In addition, although in most cases for a given venue the two values in the two columns show a same tendency, in some cases they can be considerably different. For example 'Caffe-Nero' shows a very low value of similarity between tag lists calculated with the Global and the Personalised corpus (0.14), reflecting the fact that visiting venues of that particular sub-type ('coffee-shop' chains) is rather unusual for that particular user. This demonstrates that using user mobility history to produce the weighted tag lists can significantly increase the personalisation of such representations for the specific user. A more complete data-set would derive from the aggregation of the results of a live application, when each user could retrieve his mobility history and use it to build more personalised tag-lists. This will constitute the next stage of implementation and is to be carried out as future work.

D. Similarity differences within neighbourhoods

It is also interesting to to relate these effects to different vicinities and neighbourhoods. In fact, it may be that all pubs in a certain area share some common characteristics (so defining a sub-type of 'pubs') while others in a different area can belong to a different sub-type (for example 'traditional pubs' versus 'student pubs' or 'pubs' more similar to 'bars' and 'restaurants' rather than 'clubs'). Table II computes the average similarity among venue pairs for the two categories of 'pubs' and 'coffee-shops' and for the two geographical areas chosen as example in this paper. Cosine similarity values are calculated on average between all venue pairs within the same area and belonging to different areas respectively.

	Area A1	Area A2
	Coffee shops	
Area A1	**0.1478**	*0.1076*
Area A2	*0.1076*	**0.1581**
	Pubs	
Area A1	**0.1714**	*0.1192*
Area A2	*0.1192*	0.1033

Table II
AVERAGE COSINE SIMILARITY AMONG VENUES INSIDE AND OUTSIDE TWO FIXED GEOGRAPHICAL AREA

Results in Table II show some correlation between geographical neighbourhood and average inner-outer similarity. However, whereas for the category of 'coffee-shops' the similarity among locations within the same neighbourhood is greater than between the two different neighbourhoods (thus suggesting possible different sub-types of 'coffee-shops' between the two areas) this is valid for only one neighbourhood of 'pubs' ($A1$). This could be a consequence of the fact that pubs in the more extra-central area include establishments of different characteristics, so including more traditional pubs besides others that aim to attract a different population of costumers. Further investigations are then necessary in this direction in order to include comparative studies on the personality and behavioural characteristics of the customer themselves (for example through an analysis of their mobility behaviour).

V. CONCLUSION

In this work we have proposed a methodology based on *TF-IDF* for a ranked keyword representation of venues derived from aggregation of users reviews of a number on mobile location based services. The procedure aims to return a description of the place based on the 'perception' that users (that can be of different 'type' and personality) have towards it rather than a plain objective description primarily based on venue type/category. A number of tests conducted by computing the similarity between the resulting ranked lists of keywords shows that peaks of high similarity can be found between venues of different type (but with similar overall 'atmosphere'). Self similarity between lists generated with different corpuses has also been computed in order to highlight both the most common characteristics and differences between venues of the same type (Category corpus of 'pubs' and 'coffee-shops' in our examples) or among venues included in the mobility history of a particular user (Personalised corpus). Although differences between these similarity values and those calculated using a Global corpus obtained by including all of the available venues are generally low, striking decrements in similarity values can be observed for particular venues (e.g. those with significantly different characteristics from others of a same group or those visited by a specific user). Finally the relation between geographical neighbourhood and average inner-outer similarities (within and outside a fixed area) has been examined. This could lead (in future research plans) to a re-design of the whole concept of urban-area as a characterisation based not only on the the types of places and locations in its neighbourhoods but also by the people who actually live in it.

REFERENCES

[1] Y. Tuan, *Space and place: The perspective of experience*. University of Minnesota Press, 2001.

[2] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using flickr tags to describe city cores," *Journal of Spatial Information Science*, no. 1, pp. 21–48, 2012.

[3] M. A. Vasconcelos et al., "Tips, dones and todos: uncovering user profiles in foursquare," in *Proc. of the fifth ACM international conference on Web search and data mining*, WSDM '12, pp. 653–662, 2012.

[4] A. Noulas et al., "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *The Social Mobile Web'11*, pp. –1–1, 2011.

[5] J. Cranshaw et al., "The livehoods project: Utilizing social media to understand the dynamics of a city," *Association for the Advancement of Artificial Intelligence*, 2012.

[6] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[7] J. Ghosh, *Scalable Clustering Methods for Data Mining, in Handbook of Data Mining*, ch. 10. Lawrence Erlbaum Assoc, 2003.