Cognitive Map of Tourist Behavior through Tripadvisor

Thomas Raimbault*, Gaël Chareyron*‡, Corinne Krzyzanowski-Guillot*
* *De Vinci Technology Lab – ESILV – University of Léonard de Vinci, Paris La Défense – France*
‡ *EIREST – University of Paris 1 Panthéon Sorbonne, Paris – France*
Email: {thomas.raimbault, gael.chareyron, corinne.guillot}@devinci.fr

***Abstract*—The objective of this paper is to identify, based on data from Tripadvisor, tourist behavior in how users rate a tourist place. Firstly, we propose different *correspondence analyses* (CA) on data from Tripadvisor to discover pairwise dependences between data properties (*e.g.* rating *vs.* place type, age *vs.* country). Secondly, we merge and map all our CA results as a *cognitive map*, both to bring out and understand influences between the studied concepts and to make easier human visualization.**

***Keywords*-correspondence analysis; cognitive map; tourist behavior; Tripadvisor**

I. Introduction

Tripadvisor declares itself as "the biggest travel review place on the web", today with more than 100 million reviews and more than 200 million visitors (self-proclaimed) each month. In this study, we limit the geographical area of tourism places treated as some cities in the Western Europe.

The objective of this paper is to identify, based on big data from Tripadvisor, tourist behavior in how users rate a tourism place. Ratings given by North American people are they strict? Do French people write fair reviews? How do native people feel about restaurants? *etc.*

Our contribution in this paper is twofold. On the one hand, we propose several *correspondence analyses* (CA) on data from Tripadvisor to discover pairwise dependences between different category datasets (*e.g.* rating *vs.* place type, age *vs.* country, type *vs.* country, *etc.*). CA is a statistical visualization method for picturing the associations between the levels of a two-way contingency table. On the other hand, we merge and map all our CA results as a *cognitive map*, to bring out in an human easy way all the different studied concepts, to understand their influences between them, and potentially to infer new (indirect) influences. Cognitive maps is cognitive representation model that represents concepts linked by influences, and that allows to compute propagated influences from a concept to another.

Section II presents results of the pairwise comparisons of Tripadvisor's data using correspondence analysis method. Section III shows the cognitive map resulting from our correspondence analyses. Section IV concludes this paper.

II. Correspondence analyses with R

We focus this study on ratings from Tripadvisor up to June 2014 from some places into Western Europe cities, like Barcelona, Berlin, Paris and Venice (see Table I). The different place types are: Hotel (H), Restaurant (R), and Attraction (A). Attractions may be sub-typed: organized tour ($A25$), shopping ($A26$), and bar/club ($A20$). The ratings are integers, from 1 (means very bad) to 5 (very good).

Table I: Studied (crawled) data from Tripadvisor

	Barcelona	**Berlin**	**Paris**	**Venice**
#places (id, type)	8,633	5,772	16,317	3,192
#users (id,country,age)	248,793	64,618	479,083	93,859
#ratings (1 to 5)	488,406	155,715	982,960	164,053

Correspondence analysis[1] (CA) is a multivariate statistical technique proposed by H.O. Hirschfeld [1] and later developed by the French statistician Jean-Paul Benzécri [2]. To make short, CA is a method that takes as input a "pivot table", and outputs one or more pictures of the distribution of values and variables. CA is conceptually similar to principal component analysis, but applies to qualitative rather than quantitative data. So, CA is applied to contingency tables (*i.e.* cross-tabulations). It makes little sense to compare frequencies in each cell to interpret a cross-tabulation. Then, it is essential to reduce either the rows or columns to the same base (note that row and column analyses are connected). Let us consider in Table II the matrix of *row profiles* concerning data from Barcelona about rating *versus* type of place). This matrix, computed by R with the package FactoMineR[2], gives the ratio between ratings and place's types. *E.g.* 6.8% of very bad marks (rating 1) are applied to attractions, while 25.8% to hotels and 65.0% to restaurants.

Table II: Matrix of row profiles about rating *vs.* type of place on data from Barcelona

Rating	A	A20	A25	A26	H	R	Total
1	0.068	0.011	0.011	0.002	0.258	0.650	1.000
2	0.124	0.006	0.011	0.003	0.340	0.517	1.000
3	0.173	0.005	0.009	0.004	0.388	0.422	1.000
4	0.193	0.005	0.014	0.006	0.399	0.383	1.000
5	0.317	0.006	0.051	0.011	0.289	0.327	1.000
avg profil	0.239	0.006	0.030	0.008	0.342	0.377	

The "magic" geometric technique of CA displays the matrix of row profiles in a low-dimensional space (here, 2D),

[1]The name is a translation of the French *Analyse factorielle des correspondances* (AFC), where the term *correspondance* denotes a system of associations between the elements of two sets.

[2]http://cran.r-project.org/web/packages/FactoMineR

which comes closest to the profile points. The row profiles are projected onto such a subspace for interpretation of the inter-profile positions, such that the positions of the points are consistent with their associations in the table: lower is the distance between two variables, greater is the dependency between them.

For example, Figure 1 is the result – using FactoMineR – of the CA on Table II. It follows that the best rating 5 is very commonly/probability attributed to the category of attraction (see type A very close to the value 5), but it is unlikely that the rating of an attraction was bad. The restaurant[3] rating is generally bad or medium, and an hotel is generally rated by a medium value (on Tripadvisor the average rating is 4).

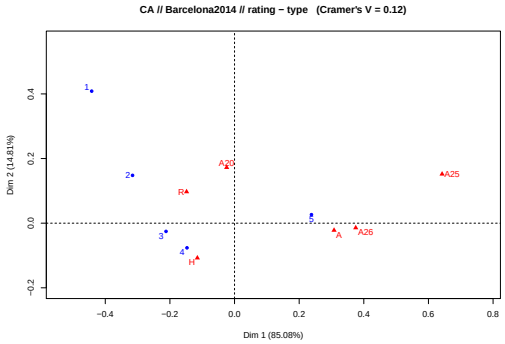


Figure 1: CA between rating and place type in Barcelona

To valid our results of statistical dependencies between variables, the *Chi-squared test*[4] (χ^2) might be a good approach, but it was not. Indeed, when the numbers are very high (like here with big data from Tripadvisor), the χ^2 almost systematically rejects the hypothesis of independence. In addition, the χ^2 can "only" determine if a dependency is significant, but does not allow to quantify the intensity of the dependency. (χ^2 varies between 0 and $+\infty$, a standardized measure is needed). For those reasons, we used the *Cramér's V measure* [3], which varies between 0 to 1 (rarely upper than 0.5 with big numbers), and which does not depend on the size of the sample. A Cramér's V value greater than 0.01 allows to conclude the dependency between variables.

All CA results about Barcelona are showed in Figure 2. One can see for example that Spanish people – in Barcelona – have a high probability to assign bad ratings (see subfigure 2a), first ratings of an user are either very good or very bad (see subfigure 2b) and mostly concern hotels (see subfigure 2e), senior people are happy with different tourist places (see subfigure 2c), native people uppermost

[3]We recall that the place type $A20$ corresponds to bars or clubs, so the same result as restaurants (R) is logical.

[4]χ^2 varies between 0 and $+\infty$, where a value close to 0 means that the hypothesis of independence between variables was accepted, while a value far from 0 means that it was that rejected (*i.e.* concludes the dependency)

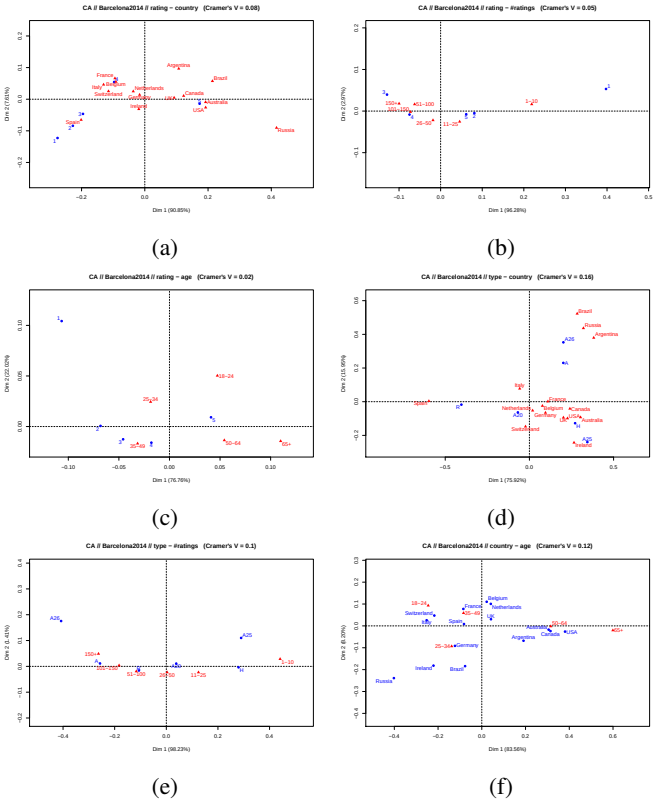


Figure 2: Different CA on Barcelona from Tripadvisor's data

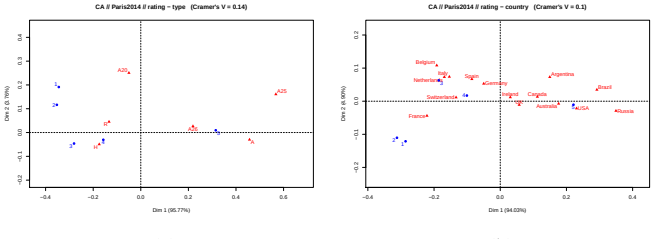
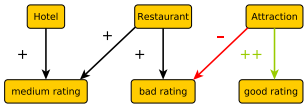


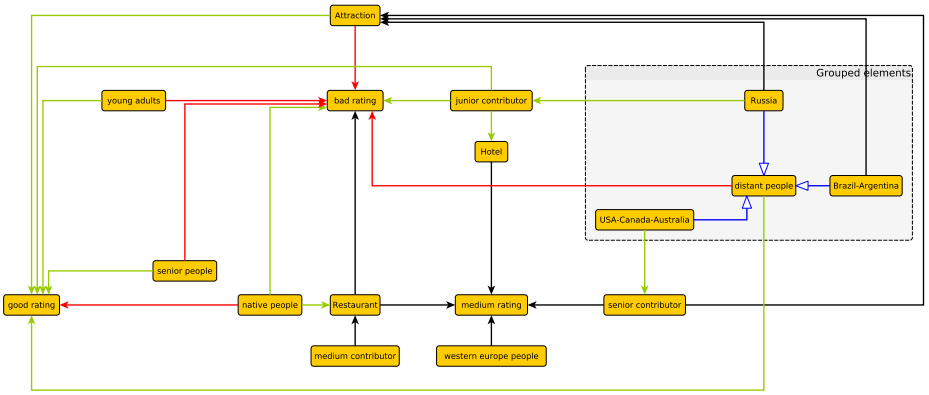
Figure 3: About Paris, CA between rating and place type (on top) and CA between rating and country (on bottom)

rate restaurants while other people rate hotels or attractions (see subfigure 2d), *etc.*

It is – very – interesting to see in this study that these results are not just about Barcelona, but these behavioral results are found in other cities of Western Europe. For instance, the comparison between place type and rating with data from Paris is showed in Figure 3a (rather than data from Barcelona in Figure 1). In this Figure 3a, it is also possible to distinguish that attractions very generally obtain good ratings, *etc.* To echo back to CA in Figure 2a one can see CA in Figure 3b, where for instance native people – French people here – generally assign bad ratings.



(a) Representation as a CM of the concept dependences showed in Figures 1 or 3a



(b) The whole cognitive map

Figure 4: The cognitive map based on our correspondence analyses about Tripadvisor's data

III. Cognitive maps

Cognitive maps (CMs) [4], also known as "mental maps", represent a set of *concepts* linked by *influences*. They are used in many fields: biology, sociology, politics... A concept is a semantic text and an influence is associated to a value; most common sets of values are $\{-,+\}$ (see [5]), the interval $[-1;1]$ (see [6]), or ordered elements, like {none, some, much, a lot} (see [7]).

In these paper, we transform the different data properties from Tripadvisor into concepts in the cognitive map model, like hotel, attraction, good rating, 35-49 (years old), Spain, *etc.* In the same way, we convert our different CA results as inference links, using the ordered value set $\{-,+,++\}$. A dependence between values is called a "positive" inference, labeled '$+$' or black colored, a high dependence is a "strong positive" influence, labeled '$++$' or green colored, and non-dependence (*i.e.* very few dependence) is a "negative influence", labeled '$-$' or red colored.

Figure 4a is the CM resulting from the CA result of dependence between place types and ratings, presented in Figures 1 and 3a: attractions are very generally rated by 5 and unlikely bad rated, while restaurants are generally rated by a medium or a bad rating. Figure 4b presents the whole CM build by the aggregation of our all previous CA results (blue arrow expresses an inheritance relation between concepts [8], *i.e.* Russian people are a subcategory of distant people according to Western Europe).

In addition to the easy human visualization feature of the cognitive map model, another important feature is that many cognitive map systems associate a causal algebra which allows to compute all propagated influences from a concept to another, according to the paths between them. Thus, new influences from indirect paths can be computed. For instance, that USA people are not especially strict (have a positive influence on medium rating); see the path USA-Canada-Australia $\xrightarrow{++}$ senior contributor $\xrightarrow{+}$ medium rating, which especially implies USA $\xrightarrow{+}$ medium rating.

IV. Conclusion and perspectives

In this paper we have presented that cognitive maps can be an adequate paradigm to semantically and schematically represent concepts from Tripadvisor and their influences between them about tourist behavior, which influences are previously computed by correspondence analyses method.

The next step is (i) to propose a study taking into account all properties from Tripadvisor's data (including gender, date of rating, *etc.*), and (ii) to infer implicit and new tourist practices, using all features of the cognitive map model. This paper finally gives a beginning (high level) way to compute tourist behavior through Tripadvisor, like to know who an user is according to how he/she rates different places.

References

[1] H. Hirschfeld, "A connection between correlation and contingency," *Cambridge Philosophical Society*, vol. 31, pp. 520–524, 1935.

[2] J.-P. Benzécri, *L'Analyse des Données. Volume II: l'Analyse des Correspondances*. Dunod, 1973.

[3] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[4] E. C. Tolman, "Cognitive maps in rats and men," *Psychological Review*, vol. 55, no. 4, pp. 189–208, 1948.

[5] R. M. Axelrod, *Structure of decision: The cognitive maps of political elites*. Princeton University Press, 1976.

[6] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1, pp. 65–75, 1986.

[7] S. Zhou, J. Zhang, and Z. Liu, "Quotient fcms – a decomposition theory for fuzzy cognitive maps," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 5, p. 593–604, 2003.

[8] A. L. Dorze, L. Chauvin, L. Garcia, D. Genest, and S. Loiseau, "Views and synthesis of cognitive maps," in *AIMSA 2012*.