2016 IEEE 30th International Conference on Advanced Information Networking and Applications

Mining and Visualizing Social Data to Inform Marketing Decisions

Jérôme Treboux, Fabian Cretton, Florian Evéquoz, Anne le Calvé and Dominique Genoud
Institute of Information Systems
University of Applied Sciences, HES-SO Valais
Sierre, Switzerland
Email: (jerome.treboux, fabian.cretton, florian.evequoz, anne.lecalve, dominique.genoud)@hevs.ch

***Abstract*—Most of today's commercial companies heavily rely on social media and community management tools to interact with their clients and analyze their online behaviour. Nonetheless, these tools still lack evolved data mining and visualization features to tailor the analysis in order to support useful marketing decisions. We present an original methodology that aims at formalizing the marketing need of the company and develop a tool that can support it. The methodology is derived from the Cross-Industry Standard Process for Data Mining (CRISP-DM) and includes additional steps dedicated to the design and development of visualizations of mined data. We followed the methodology in two use cases with Swiss companies. First, we developed a prototype that aims at understanding the needs of tourists based on Flickr and Instagram data. In that use case, we extend the existing literature by enriching hashtags analysis methods with a semantic network based on Linked Data. Second, we analyzed internal customer data of an online discount retailer to help them define guerilla marketing measures. We report on the challenges of integrating Facebook data in the process. Informal feedback from domain experts confirms the strong potential of such advanced analytic features based on social data to inform marketing decisions.**

***Keywords*-data mining; data visualization; social media; linked data; marketing; methodology; knime; d3**

I. Introduction

Nowadays, social media are increasingly used as an information source for companies to better profile and serve their clients. The quantity of information available from social media grows significantly every year. For example, in 2015 there are 222 Million more active social media account compared to 2014 (+12%). This reprensent 2.078 Billion active social media accounts (29% of the total population). Some studies estimate that in average a user may spend as much as 16 minutes per hour every day on social media and that 23% of Facebook users login at least 5 times a day[1][2]. Because of this explosion, social media became a very important data source for companies to develop and adapt their business in order to meet customer needs and attract new customers. Indeed, user-generated content is large - around 500 Million tweets posted and 1.8 Billion pictures posted every day on social media[3]. Due to this growing quantity of freely expressed opinions available a whole new fast growing sector of activity is emerging and many software tools propose ways to handle customers activities on social media. Those community management tools offer social media analytics that allow the company to assess how the users perceive the quality of its services, how much visible the company is, and how much people engage with its pages. Information about the activity and visibility of the company over networks, but also socio-demographic information and sentiment analysis is provided, often in a visual manner. However important features are still missing from most commercial tools : generic features (multilingual text analysis, fusion of social data with company-internal data e.g. customer data), and of course features specific to a particular domain of activity.

On the other hand, information visualization has emerged as a way to get insight from large quantities of data. Information visualization has been defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition"[4]. In general, information visualization provides helpful techniques to make sense of complex data, providing the ability to see the broad context of the data or focus on specific details.

In this work, we explore how a combination of social data mining and information visualization can inform marketing decisions. We present the data mining tools and prototype visualizations that we developed for marketing specialists in two different use cases: a swiss regional tourism promotion agency (Valais Wallis Promotion) and a swiss online discount retailer (Qoqa.ch). We shortly introduce here some of the specific challenges of both domains that we deem as relevant in a marketing context.

In the touristic domain, the quality of service is of crucial importance, therefore having a way to gather user feedback about the service is key. Nowadays, surveys are the most used way to assess tourists' opinions. Properly creating and analysing the results of surveys is very time consuming and expensive and requires specific skills that traditional actors in tourism lack. As a result, surveys are done generally by government agencies for a whole area regrouping a lot of destinations and services. The aggregated result is of little help to specific actors who want to use this data to improve their strategy and marketing. Finally, as the timespan between the start of the survey and the final result could be very long, it is difficult to adapt in a timely manner. Relying on social data therefore not only provides an opportunity to get alternative information, but

1550-445X/16 $31.00 © 2016 IEEE
DOI 10.1109/AINA.2016.130

also to speed up the whole process as the data is available immediately.

In the domain of online commercial retailing, an important challenge is to get insights about specific characteristics of domain-related company-internal data, for which no generic data transformation and visualization tools exists. Another challenge is to merge social data with company-internal data, in order to identify the behaviour and feedback of customers on social websites, and differentiate them from other people's data.

Our work tackles those specific challenges in two use cases. Overall, our work consists in (1) the creation of a simple methodology to design visual analytics solutions, based on a combination of data mining and visualization design approach, (2) the mining of social data, in particular textual and geocoding data relevant for each use case and (3) the design and development of prototype visualizations based on the gathered data. Moreover, informal qualitative feedback from domain experts (marketing specialists) was collected. They comment on the utility and usefulness of the prototypes.

The paper is organized as follows: in section 2, we present our methodology. In section 3, we describe two use cases with details about the methodology implementation, and all steps from data extraction to the implemented visualization. We then present following discussion and users feedbacks in section 4. In section 5, we review previous related works and finally conclude in section 6.

II. Methodology

The design and development methodology we defined for this project is based on CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM describes a process commonly-used by data mining experts in the industry [5]. We adapted it slightly, in particular we included specific steps dedicated to the design and development of the visualizations of mined data. The methodology (see Figure I) starts with the identification of the client needs through interviews. We drew inspiration from the 7-step approach described in [6] to identify a good business case. This approach consists in the following steps:

- Identify the Actors
- Identify the Goal
- Define the Pre-Conditions
- Define the Post-Conditions
- Describe the Main Flow
- Describe the Exceptions
- Describe the Alternate Flow

The outcomes of this approach are then materialized in a concrete use case.

Then, two branches occur in parallel: the data mining branch and the visualization branch. In the data mining branch, like prescribed by CRISP-DM, work starts with data extraction and continues with data understanding. In the visualization branch, the goals definition, derived from the use case, drive the design of the visualization. Based on this design and the data extracted, data can be processed (e.g. cleaning, formatting, aggregating . . .) in order to match the need of the visualization. Once this is done, the final visualization can be developed.

In the next section, we present our two use cases, structured according to the steps of the methodology presented above.

III. Use cases

A. *"Valais Wallis Promotion"*

Valais Wallis Promotion (VWP) is a Swiss company who does regional tourism promotion, regrouping all resorts and places in the canton Valais.

1) Use Case Definition: The initial need formulated by VWP was to understand what people really enjoy and look for in a given touristic destination, in order to compare the real tourists point of view with the marketing done by that destination. As an example, this tool would allow to spot out that advertising is done about a particular sport but the tourists coming there mostly talk about another sport or another topic (as food). The marketing experts can start to investigate the situation, and determine how the advertisement needs to be tailored or re-oriented to achieve a successful communication.

To formalize this need, we applied the methodology described in [6]. Because of the exploratory nature of the project, only 4 of the 7 steps were done. First, the actors identification: tourists and tourism offices. Then, the goal: expectations of tourists visiting in a specific destination. The pre-conditions: data is published on existing Social Media platforms, pictures and comments are freely accessible. Lastly, the post-conditions: be able to identify the clients needs and adapt the marketing accordingly.

The research question concerning this business case can be summarized as: "Based on social media data and pictures published online by tourists, how can the tourist offices identify the clients expectations in order to adapt their marketing strategy accordingly".

2) Visualization goals: VWP needs to explore data and understand what people are saying about a specific destination. Therefore the visualization must present in one shot a big amount of browsable data, as a supporting tool to adapt the marketing strategy.

3) Data extraction: We identified Instagram and Flickr as two relevant data sources. Their open API is well suited for our data extraction. A specific development was done for each API, but they both provide the needed functionalities in a similar way. The API works as a search method called with keywords. We defined a list of destinations and performed one search query for each one of them, getting a set of results with all pictures containing the name of the destination in the description. We then retrieved pictures and their comments,

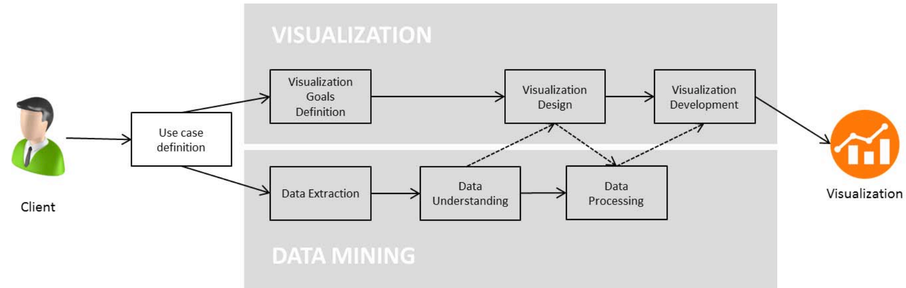

Figure 1. The methodology used to create the visualization in our use cases is based on CRISP-DM, with the addition of a visualization branch.

the location when provided, as well as the number of likes and commentaries. This information was further completed with the corresponding user's data, as for instance the origin of the user. The whole data set was then stored in our local relational database.

4) Data understanding: The purpose of data understanding was to validate the extracted corpus quantitatively and qualitatively. We were well-aware that our extraction method based on a predefined list of destinations names had two weaknesses, the two classical false negatives and false positives. False negatives are the pictures indeed about the concerned destinations but that could not be identified because the place name was not explicitly written or was misspelled. On the other hand, false positives are places from other countries but with the same name, as for instance Sion, the capital of Valais, which does correspond to at least 5 other places around the world.

To disambiguate the homonyms, we did validate the pictures with a geolocalization, controlling that their geolocation correspond to the region of Valais. Ending-up with a clean corpus of 147'900 unique pictures, we choose not to worry about false negatives as our input data was large enough for our prototype.

5) Visualization design: The main question of our business case is "Identify the tourist needs to adapt the Marketing". To address it, we chose to visualize two different aspects of the data: (1) the origins and destinations of tourists, to help identify general trends and (2) the topics of interests of tourists, based on what they mention on social media.

To achieve this, we designed three visualizations. The first two ones address the challenge of visualizing origins and destinations. First, a dot map shows the overall distribution around the world of tourists coming to Valais based on the available data. Second, a proportional area table allows us to see at a glance how many pictures tourists coming from a specific country of origin have taken at a specific destination in Valais. The last visualization addresses the question of tourists' topics of interests. As topics extracted from textual social data will be enriched with semantic concepts, we decided to use a tree-like visualization grouping each topic into higher-level concepts, namely with a sunburst visualization. The visualizations are presented in greater details in section "Visualization development" below.

6) Data processing: A widespread habit of social media users can be turned into an asset: the micro-blogging trend to represent an important term with a hashtag allows easier keywords extraction from the texts[7][8][9]. As demonstrated in our state of the art, we wanted to push the research in that domain by applying multilingual semantics to improve the results. The goal of the semantic layer is to move from a flat list of words (the tags), in different languages, to a knowledge graph. This knowledge graph is a multilingual network of concepts and their relationships. For this specific use case we did focus on three languages: English, French and German. By handling synonyms and translations we are able to relate different tags to a single concept (for instance a bicycle for the tags "bicycle", "bike", "vélo", "Fahrrad"). By handling hyponyms, we can link more general and more specific concepts together, thus knowing that "Mountain bike" (hyperonym-more specific) is also a tag about the "bicycle" concept (hypernym-more general), and that talking about "cycling" (hyperonym) is also talking about "sport" (hypernym). Instead of working with simple vocabularies or thesaurus, we decided to use ontologies which are a powerful way to model a domain with all its concepts and their relationships. Moreover they are the building block of the Linked Data technology [10] [11] that allows facilitating the access of the vocabularies by making them directly available on the web, but also to explicitly link different data sources and thus respond to our need of a cross-language and cross-domain reference vocabulary.

As a first step, we had to make a choice amongst the available multilingual ontologies that relates to linguistic, and analyzed the one provided by the Linguistic Linked Open Data initiative [12]. Taking into consideration our specific needs, we identified BabelNet [13], freely available for research purpose, as the perfect candidate for us. BabelNet being part of the Linked Data Cloud [14], information about a concept can be further completed by other data sources, as for instance the Linked Data version of Wikipedia: DBPedia [15]. Our intend was to link the hashtags to BabelNet concepts, and then complement the knowledge with the WikiPedia/DBPedia categories for the purpose of clustering and classification.

Our methodology to create a multilingual network of words on top of the flat list of tags was to query the ontology for each tag to find a corresponding entry based on a string match. BabelNet can be queried in different ways: HTTP requests, java API or a SPARQL end-point (SPARQL being the query language for Linked Data). For our first run, we choose to use SPARQL to find a correspondence for tags used more than once, 30'000 in total. For each tag, a query was done to find a lexical entry for any of the three languages, which resulted in a total of 13'000 tags successfully linked to a concept. For each identified entry, translations and first level hyponyms where queried as well to generate the graph. Having more than 40% of the tags corresponding to a concept, which was a first positive result for us, we decided not to investigate further in tags decomposition technics to handle tags made of more than one word. With our queries, some composed words as "lake Geneva" already found the correct mapping without further treatment. Another well know difficulty in this kind of process is the handling of polysemy and disambiguation that we also choose to keep for further work. Indeed, our tool is a help to group the tags and create a hierarchy, but it is the user who will take the final decisions and evaluate if the presented information is relevant or not. It is to be noted that our choice for BabelNet was supported by the fact that the author also provide Babelfy [16], an online service that gives state-of-the-art performances for word sense disambiguation and that will be useful for our further handling of homonyms. About the resulting graph, here are a few examples of tags that were regrouped as a single concept: "switzerland; schweiz; suisse; svizzera; swizerland", "clouds; nuages; cloudy; wolken", "lac_leman; lake_geneva; genfersee", "bike; velo; bicycle; rad; bikes; wheel; bicicletta". We believe this work is an added value and would have been much more difficult to obtain with another approach.

To fulfill the use case, the need for clustering and classification of the concepts was solved by using two available hierarchies in our knowledge base: the hyponymy hierarchy and the Wikipedia categories hierarchy. Moving up such a hierarchy allows finding generic concepts or themes. In the hyponymy hierarchy, 3'000 concepts were identified as hypernyms, having at least one hyponym. The bike concept is correctly the hyperonym of tags "vtt; mountainbike". Sport directly regroup: "cyclisme; cycling; biking; cyclosport; course; racing; race; rowing; skiing; sking; skiers; skieurs; ski; skier; skifahren; riding; équitation; sledding; luge; gymnastics; gym; leader; second; rockclimbing; escalade; klettern; athletics; skating". And querying more deeply the hierarchy all different sports are listed as "icehockey; eishockey; running; track; jumping; springen; skateboarding; skateboarder; iceskating; roller; rollerblading; inline; foot-

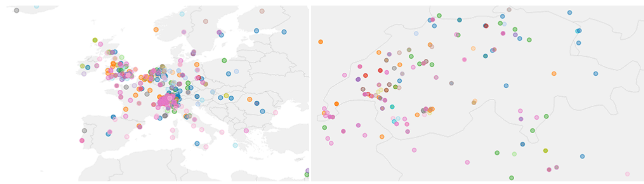

Figure 2. Origin of tourists visiting Valais according to Flickr. Position of dots correspond to origin of tourist. Color corresponds to a destination. European map (left), Swiss map (right).

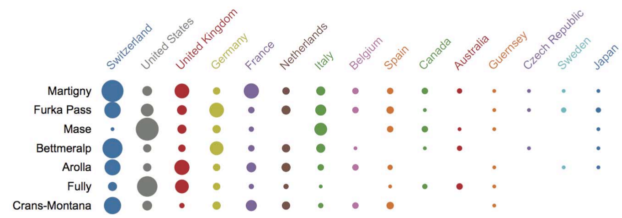

Figure 3. Extract of a tabular visualization of pictures from Valais posted on Flickr. Each bubble represents the pictures taken at a specific destination (row) by people coming from a specific country of origin (column and color). The area of the bubble is proportional to number of pictures.

ball; soccer; fußball; wrestling; boxing;.etc.". Concerning the Wikipedia categories, a queryable semantic representation available with DBPedia allows to bring together "moutain bike" and "bobsleigh" under the "Racing vehicles" category, and then the more general "racing" category. But it also groups "moutain bike" and "bicycle" under "cycle types". All the given examples are tags from social media actually found in our use case.

7) Visualization Development:

Visualization of Origin and Destination*:* The goal of the visualizations described in this paragraph are to overview the origin and destination of tourists based on pictures published on Flickr. The pictures are geo-coded, therefore each of them can be mapped to a specific destination in Valais. Besides, Flickr users publish their origin in their profile, which is also available through the API. We have designed and developed two origin-destination representations by combining users provenance with their pictures location.

The first one is a world map where each picture taken in Valais is represented as a dot on the user's origin. For this map visible in Figure 2, 33 destinations are analyzed. The color of the dot corresponds to a specific destination, the place where the picture was taken. As an example, for each picture of Martigny, a red dot is displayed on each user's original location in the whole world. The map can be zoomed to display with precision a country or city, showing for instance all the pictures taken in Martigny from travelers living in England or more precisely in London. This tool brings insight about the origins of travelers (among those who post their pictures on Flickr and disclose their origin). To answer more specific destination-oriented questions we worked on the proportional area table shown in Figure 3.

This table clearly highlights how the places (in rows) are more or less photographed by citizens of different countries (in columns). The size of each bubble is proportional to the quantity of pictures corresponding to a given origin and destination. Moreover, sorting options can be used to put in light different patterns, by regrouping provenance countries by continent or playing with latitude and longitude for example.

Those two representations are useful to identify origin-based group of clients, and then adapt specific marketing strategies according to their behavior and preferences. The first visualization reveals for instance that Flick users who publish pictures about Valais are mostly Europeans, whereas the second one points out that Japanese tourists publish only about one fifth of the destinations we are analyzing.

Visualization of Semantic Data: Our resulting semantic graph is an abstract construct that is hard to render visually in order to identify the tourists needs. To determine the importance of each concepts and give it a weight, we summed up the count of the corresponding tags (synonyms and translations) plus the count of all more specific concepts (hyponyms). The tags hierarchy of the four nodes with the highest count was displayed in dynamic and zoomable sunburst.

Sunbursts are based on pie chart to represent hierarchical data, depicted by concentric circles. They allow the exploration of hierarchical clusters of nodes. The hyponymy hierarchy can be represented as the center of the circle (See Figure 4). Therefore we visualized them as a concentric circles that shows the sequence of semantic clustering steps of the tags found in the original textual data. As we can see on that figure there are five hierarchical levels of clustering of the initial set of tags found in the data.

The user interface we developed lets our client use a search field and look for a specific concept. For example, let's say the tourism office wants to promote skiing. Searching for "*ski*" shows a concentric circle with all linked concept as "*slalom, schuss, etc.*. The weight given to each of these word gives an idea of its popularity amongst tourists comments. The office is then able to adapt their communication and replace the "*ski*" concept by "*slalom*" to better match the client's expectations.

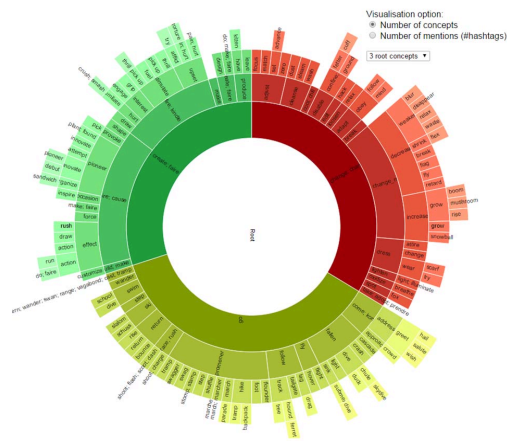

Figure 4. Sunburst visualization presenting the concepts most mentioned by tourists in the picture caption posted on their Instagram profile.

B. Use Case "Qoqa"

Qoqa is a very successful swiss online discount retailer who is offering to buy a unique different article everyday, in a limited amount, with a high discount.

1) Use Case Definition: This use case is based on the company's own data combined with information from Facebook. The original idea was to identify the Facebook followers who are not customer yet, in order to motivate them to make their first purchase. But while applying the methodology and defining the pre-conditions, we got a first negative result due to a restrictive update in the Facebook API and their Open Graph: the full followers list of an account is no longer accessible. We tried to use an analytic tool called Engagor and used internally. Engagor is a powerful online community management tool, providing the possibility of downloading a user list, but only of users who are active on the page. We were thus compelled to adapt the goal. After discussions with the client, we chose to focus our analyzis on the internal customers database containing name, first name and postal code only. The new goal is to define where a Guerilla Marketing should be done to reach the right customer segment, based on some localized profiling of the customers.

We formalized the need with the proposed methodology. First, the actors identification: Qoqa and its clients. Then, the goal: improve the marketing operations to better target the customers. The pre-conditions: data is only available in the local database. Lastly, the post-conditions: identify the clients profiles and locations in order to improve the localized marketing.

We ended up with the following research question: how to identify the most strategic locations for specific marketing actions ?

2) Visualization goals: After multiple iterations with the client we defined the goal of the project's visualization. They wanted to see a detailed map of the Swiss customers, aggregated at the two levels of cantons and districts. This visualization must allow them to visually identify the most important regions according to the number of customers. Moreover proportion of men and women must also be represented.

3) Data extraction: To create this representation, we used data from the customers database of Qoqa, after having agreed on confidentiality issues.

4) Data understanding: The client database was easy to understand as the extract contained only the family name, firstname, and the city postal code. There was no field about

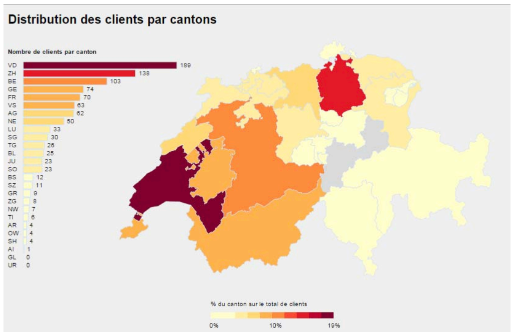

Figure 5. Extraction of Qoqa consumer distribution by canton. The colors represent the importance of a canton compared to others.

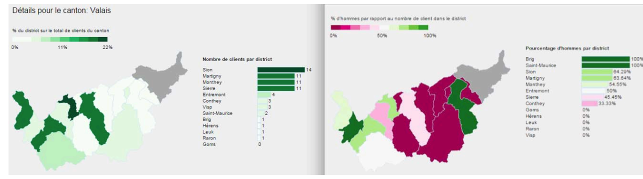

Figure 6. Example of client (left) and gender (right) distribution by district in the canton of Valais.

the gender, and that would be our job to guess it.

5) Visualization design: Qoqa has explicitely asked to visualize the data in maps. So we decided to use choropleth maps, with colors associated to specific regions, to represent the amount of customers per region. However, choropleth maps are not well suited to precise comparisons and may be misleading when comparing amount of customers (e.g. the size of a region directly impacts how much importance its color seem to have). Therefore, we chose to complement the maps with simple ordered bar charts presenting the same data and that would allow a more precise comparisons of regions. Those visualizations are presented in details in section "Visualization development" below.

6) Data processing: To fit to the visualization design some data processing steps were mandatory to localize the clients on a map and also handle a gender filter.

To link the postal codes to cantons and districts, the API of Geonames.org was used. A match was performed on the postal codes, allowing us to get latitude and longitude coordinates needed for our visualization tool. Furthermore, links between postal code and districts were also retrieved for further use.

To find out the gender, a text-mining process was run to differenciate masculine and feminine given names. This text analysis was done through a data-mining software called KNIME. The whole process is based on a dictionary for a given language. Some records were deleted (with the client's consent) to solve the issue of unisex names (firstname that can be used by man or woman) or some unclassified names (e.g. "M. & Mrs. Smith").

7) Visualization Development: The final visualization developed for the client provides a one page overview of the consumer distribution around the country (See Figure 5). The intensity of the color of a canton maps to the proportion of the total amount of customers who come from that canton. The bar chart next to the map shows the absolute number of customers (Note: for confidentiality reasons, the data presented in the visualizations are fake).

We also developed a detailed view of each canton. This detailed view has two parts. The first part gives a detailed customers distribution by district, based on the global number of customers in the canton (See Figure 6). The visual presentation of the data is the same as before, but on a different granularity level.

The second part of the detailed view presents the gender distribution by district. We used a diverging color palette [17] to show the relative predominance of male or female customers in each district. This visualization is very useful in terms of Guerilla Marketing, to promote the right product in the right area (See Figure 6).

IV. Discussion & Users feedbacks

Tourism experts from the Observatoire Valaisan du Tourisme (OVT) a government-funded agency that monitors tourism in the canton of Valais, evaluate our data representation for the first use case (Valais Wallis Promotion) as ”very useful”. Playing freely with our visualizations tools, they commented that the current marketing strategy ”mainly based on socio-demographic profiles” could benefit from our approach, and they see our demonstrators as ”a means to rethink marketing and obtain new information in order to optimize the different client segments”. They point out that our tool is currently ”more suited to applied research” and would need some adaptation to answer the specific needs of actors in the tourism industry. They confirmed our sentiment that ”information from social media is underused” as the branch professionals rely on social networks mostly ”to communicate” in order to attract tourists attention. However the project gave them some insight to ”better understand social networks and increase [their] comprehension of tourists needs” regarding ”specific social networks”. For future research, in order to present the visualizations to tourism actors, they mention the challenge of ”simplifying [the visualizations] without loss of information”. To improve the data understanding process, they propose to include the origins of users in order to take into account some cultural aspects. Finally, for a marketing perspective, they see ”socio-demographic and semantic tools” as a powerful solution to optimize the messages with ”the adequate language” for the different types of host.

V. Related work

From the early days of the Web 2.0, social network data has been seen as extremely valuable to support marketing decisions, and the study presented in [18] proved word of mouth to be a major player regarding a company's success. At this time, different models have been proposed [19][20] to measure the network value of customers and enhance viral marketing. The need to better understand the content of the customers' text lead to the improvement of the algorithms to process natural language (NLP). Opinion Mining and Sentiment Analysis became key tools to evaluate the opinion of the users[21]. The power of those techniques was successfully applied to real-time data for predictive analysis [22], outperforming the gold-standard information market in the presented use-case of the movies box-office revenues.

Regarding hashtags analysis, considering hashtags as keywords defined by the author himself and thus relevant for us to further understand its text, interesting research must be mentioned. Classical NLP and machine learning techniques have been applied to cluster the tags and get a better classification. The relevance of the approach is demonstrated by Muntean[23] and Antenucci[24] who also point out that the result suffer from the lack of semantics. Moving from text handling to higher level of knowledge handling[25] gives the ability to manipulate words instead of simple character strings, thus dealing effectively with synonyms and other semantically related terms or concepts with the help of a thesaurus or a richer ontology. The result of the machine learning classification task are improved in Pöschko[26] and Djuana[27] by mapping the tags to the English WordNet Ontology[28]. We took their work as a starting point for our data understanding and data processing steps, choosing a multilingual and general ontology to deal with the different languages and domains of our rich texts corpus.

Visualization of social data has received an increased interest from researchers and practitioners in recent years. We review some work that has links to ours in terms of data source and visualizations employed. [29] uses dot maps to represent the results of filtering queries over large sets of Flickr pictures. Several attributes of the picture can be used as query parameter in their system, e.g. its location, the name of the photographer, the time when it was taken, etc. The goal of their tool is to find the most relevant photographs in order to provide a summary answer to a given query. [30] uses photos posted on Instagram to try to understand people's activities in cities. Spatio-temporal visualizations show thumbnails of pictures taken in different cities arranged according to several parameters, including time and hue. They use image plot visualization (montages) to analyze how many pictures were taken at given time in specific touristic destinations. The visualizations produced are elegant, but not meant primarily for analysis. [31] leverages Instagram pictures to fuel a real-time monitoring system that detects events based on pictures posted on the social network. As a "web-based ambient visualization meant to be shown on large displays", e.g. in newsrooms for journalists, its interface presents simple statistics about data in the form of line graphs and map of activity, plus a stream of relevant pictures of events detected by the system. Finally, [32] analyses sentiments based on Twitter data. Visualizations employed include a dot map and a timeline view of tweets presented as colored dots.

VI. Conclusion

In this paper we presented the methodology and techniques we applied to mine and visualize social media data in order to support marketing decisions for two Swiss cases. To get richer results than community management tools and their generic-purposed analytic features, we used a use-case specific approach to better serve the particular needs. We applied a slightly adapted CRISP-DM methodology and carried out the steps of data extraction, data understanding, visualization design, data processing and visualization. For social data extraction we have seen that social medias platform offer a variety of API to programatically query the content, but we are bound to limitations and features that vary over time. A recent change in the Facebook API making the followers list of an account inaccessible forced us to adapt one of the use case. We focused on the company's customers data performing localized profiling and gender determination to help identifying where a Guerilla Marketing should be done. In the other use case, we tackled the problem of understanding the tourists expectations in order to adapt the marketing strategy. We presented a novel way to create a multilingual network of words and concepts on top of the hashtags keywords found in users' comments. This semantic knowledge representation was assessed as very valuable by tourist experts. They confirmed that the information from social media is really under-used, at least in Switzerland, and could effectively support marketing decisions in their opinion.

Acknowledgment

We would like to thank Ontotext who supported us with a research license for OWLIM. We also thank Roberto Navigli and colleagues (Sapienza University of Rome) for providing the research license for BabelNet.

References

[1] J. Casas, "Social media trends 2014: The latest data, infographics and statistics," 2015. [Online]. Available: http://postcron.com/en/blog/social-media-trends-2014/

[2] S. Kemp, "Digital, social & mobile worldwide in 2015," jan 2015. [Online]. Available: http://wearesocial.com/uk/special-reports/digital-social-mobile-worldwide-2015

[3] J. Oldham, "Socially connected: How brand, retail and media websites are evolving," mar 2015. [Online]. Available: http://www.engagesciences.com/socially-connected-brand-websites-evolving/

[4] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[5] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[6] I. Consulting, "7 steps to writing business use cases," may 2015. [Online]. Available: http://www.iag.biz/white-papers-templates/business-analysis-resources/7-steps-to-writing-business-use-cases.html

[7] D. Romero, B. Meader, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *proceedings of the 20th International conference on World wide web*. ACM, 2011, pp. 695–704.

[8] E. Tonkin, H. D. Pfeiffer, and G. Tourte, "Twitter, information sharing and the london riots?" in *Bulletin of the American Society for Information Science and Technology*, 2012, pp. 49–57.

[9] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, August 2010, pp. 241–249.

[10] T. Berners-Lee, "Linked data. design issues for the world wide web," *World Wide Web Consortium. http://www. w3. org/DesignIssues/LinkedData. html*, 2006.

[11] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2009.

[12] C. Chiarcos, "Linguistic linked open data (llod)–building the cloud," in *Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction*. Citeseer, 2013, p. 1.

[13] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[14] M. Ehrmann, F. Cecconi, D. Vannella, J. Mccrae, P. Cimiano, and R. Navigli, "Representing multilingual data as linked data: the case of babelnet 2.0," in *Proc. of LREC*, vol. 14, 2014, pp. 401–408.

[15] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[16] A. Moro, F. Cecconi, and R. Navigli, "Multilingual word sense disambiguation and entity linking for everybody," *Proc. of ISWC (P&D)*, pp. 25–28, 2014.

[17] M. Harrower and C. A. Brewer, "Colorbrewer.org: An online tool for selecting color schemes for maps. the cartographic," *Journal*, pp. 27–37, 2003.

[18] F. F. Reichheld, "The one number you need to grow," *Harvard business review*, vol. 81, no. 12, pp. 46–55, 2003.

[19] P. Domingos, "Mining social networks for viral marketing," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80–82, 2005.

[20] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.

[21] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[22] S. Asur, B. Huberman *et al.*, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.

[23] C. I. Muntean, G. A. Morar, and D. Moldovan, "Exploring the meaning behind twitter hashtags through clustering," in *Business Information Systems Workshops*. Springer Berlin Heidelberg, January 2012, pp. 231–242.

[24] A. M. D. Antenucci, G. Handy and M. Tinkerhess, "Classification of tweets via clustering of hashtags," 2011.

[25] P. Buitelaar and P. Cimiano, *Ontology learning and population: bridging the gap between text and knowledge*. Ios Press, 2008, vol. 167.

[26] J. Poschko, "Exploring twitter hashtags," in *arXiv preprint arXiv:1111.6553*, 2011.

[27] E. Djuana, Y. Xu, and Y. Li, "Constructing tag ontology from folksonomy based on wordnet," in *proceedings of the IADIS International Conference on Internet Technologies and Society 2011, International Association for Development of the Information Society (IADIS)*, 2011.

[28] C. Fellbaum, "Wordnet: An electronic lexical database," in *Cambridge, MA: MIT Press*, 1998, p. 423.

[29] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 2006, pp. 89–98.

[30] N. Hochman and L. Manovich, "Zooming into an instagram city: Reading the local through social media," *First Monday*, vol. 18, no. 7, 2013. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/4711

[31] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman, "Citybeat: Real-time social media visualization of hyper-local city data," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion, 2014, pp. 167–170.

[32] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. Keim, L.-E. Haug, M.-C. Hsu *et al.*, "Visual sentiment analysis on twitter data streams," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 277–278.