Mining User Queries with Information Extraction Methods and Linked Data

*

Université libre de Bruxelles (ULB)
ReSIC Research Center
Information and Communication Science Department
Avenue F.D. Roosevelt 50 – CP 123 – B-1050 Brussels, Belgium
{}@ulb.ac.be

Abstract

Purpose Although the continuous development of Web analytics tools offers new perspectives for heritage institutions to improve their understanding of user needs and behaviour, their functionalities remain relatively limited to analyse how end users interact with complex online catalogue. However, advanced usage of Web analytics tools allows to capture the content of user queries, which are more often than not deleted from log files of the catalog database after a certain period. Despite their relevant nature, the manual analysis of large volumes of user queries is problematic. This paper demonstrates the possibilities and limits of using information extraction techniques to gather a better understanding of the nature of user queries in a fully automated manner. In order to do so, the paper presents a large-scale case-study conducted at the Royal Library of Belgium consisting of a data set of 83.854 queries resulting from 29.812 visits over a 12 month period of the historical newspapers platform BelgicaPress. By making use of information extraction methods, knowledge bases and various authority files, the paper present the possibilities and limits to assign queries to pre-defined categories.

Design/methodology/approach

Findings

Originality/value

Keywords User query; Query Classification; Digital Libraries; Cultural Heritage

Paper type Case study

1 Introduction

Both policy makers and the public are increasingly regarding libraries, archives and museums as content and service providers who operate in the same market (and compete for the same customers) as commercial information providers. This situation is reflected

*Corresponding author

in the adoption within the cultural heritage sector of the common definition of the "quality" of information systems and services by ISO, which focusses on the "fitness for purpose" (ISO, 2005). This interpretation of quality refers to the idea of self-regulating markets where demand directly influences supply as users/consumers are empowered to decide what information is of use. The work of Vesa Suominen can be consulted for a more in-depth discussion on the problem and limits of "userism" within the context of libraries (Suominen, 2007).

Within this context, cultural heritage institutions have been eagerly making use of Web analytics tools to quantify the interaction between their collections and end-users. The shiny dashboards of popular tools such as Google Analytics do provide useful features to understand how many end-users interact with a website, from where they come, how long they stay or with which specific Web pages they interact. However, this approach does not provide a detailed analysis of how patrons interact for example with search forms. In the context of the MADDLAIN project, the functionalities of the open-source Web analytics tool Piwik were extended by implementing custom scripts allowing to capture relevant information contained within the URL's of the user's browser. This method resulted in the creation of a massive data set, leading to the problem of how to make sense out of such a large body of usage data. By giving a concrete overview of the possibilities and limits of information extraction methods, this paper wishes to give a more nuanced understanding of the complexity of interpreting large volumes of usage data and how to make sense of them in an automated manner.

The paper starts with a short literature overview of relevant research on the aggregation and interpretation of usage data in the cultural heritage sector and the application of Information Extraction Methods, after which the case study from the Royal Library in Belgium and the related research questions is presented in detail.

2 Related work

Online user behaviour has been increasingly analysed in the cultural heritage field, especially since the launch of Google Analytics in 2005. As highlighted by Kelly (2014), Web Analytics can lead to developing enhancements to the architecture, metadata or content of a digital library to improve both user experience and success. Various methodologies and metrics have been developed to fit archive and library website specificities. For example, Fagan (2014) illustrated how commercial key performance indicators can be adapted to an academic library environment. Still, beyond assessment tools helping to collect user experience, "additional tools for automating and analyzing this data are still needed to make it a widespread practice [for archives]" (Kelly, 2017).

Thus, as underlined by Zavalina and Vassilieva (2014), very few studies published by large scale digital libraries examine the content of the user search queries. Ceccarelli, Gordea, Lucchese, Nardini and Tolomei (2011) used Europeana query logs, but more as means for developing assistance functionalities such as a query recommender system than as objects of study per se. Likewise, Dijkshoorn et al. (2014) considered the log files from the Rijksmuseum as an help to combine user queries with external vocabularies published as Linked Data, in the attempt to diversify search results. In both cases, no text mining methods appear to have been performed on user queries. By contrast, Zavalina (2007) mentioned, in the context of the IMLS Digital Collection query logs, that some processing of the queries (truncating plural, excluding stopwords such as prepositions,

etc.) has been done before categorizing and finding semantic matches with a controlled vocabulary. However, the whole process, including the extraction of all query strings from the log files, was done manually on a corpus containing less tan 1 000 queries. This example highlights the potential of computational methods in this context: using a script to semi-automatically process the data and automatically extract information can save time and could be applied to larger datasets. In our previous work (Chardonnens and Hengchen, 2017), we started to explore that potential by describing in detail a 5-step method to perform text mining on a large volume of log files from the State Archives of Belgium. This paper aims to go further by including information extraction methods.

In the more specific area of online digitised newspaper, De Wilde and Hengchen (2016) presented a case study from the *Historische Kranten* project. Before focusing on the potential of named entity recognition and linked data to enrich multilingual archives metadata, they evaluated user demand. To do so, they tracked individual queries over a 4-year period. Their findings revealed that, according to the ten most popular keywords, locations are especially favored. Although promising, the analysis is not developed.

Eventually, Gooding (2016) performed an overall analysis of the information behaviour of users of Welsh Newspapers Online Website. Using in a complementary way the possibilities offered by Google Analytics and web server logs, he observes than the first one is not tailored for academic research and provides a weaker source for in-depth analysis, due to the opacity of data processing and the impossibility for the user to export raw data. The web logs allow him to identify most viewed newspaper titles, most viewed decades and most commonly viewed page numbers. At last, he observed than "over half of pageviews are dedicated to interacting with the web interface rather than than the historical sources". While filling a gap in the literature, he offers a paper which leaves room for experimentation in the analysis of the content of user queries themselves.

Progress can be made by building on experience gained in the broader field of web query classification. Well-known challenges for query labelling are caused by the nature of web query, wich are usually short in length, ambiguous and noisy (wrong spelling).
-NER
-Query classification / categ. Beitzel (2007) ; Demartini (2016) ; Khoury 2009 + 2011 -> Limites Wikipedia
- Ranking entity types Demartini (2013)
- Desambig + context - expand queries 'Previous studies have confirmed the importance of search context in QC. (Cao et al., 2009) considered the context to be both previous queries within the same session and pages of the clicked urls.' ; Le et al. (2011) ; Le and Bernardi (2012)

3 Research question and methodology

This article wishes to demonstrate how automated methods can help to answer a specific research question when facing a large corpus of user queries. Within the context of the evaluation of its services, the Royal Belgian Library wishes to understand the information needs of its patrons in regards to the historical newspapers published as *BelgicaPress*.[1] Launched in 2015, this search engine provides online access to more than two million pages of digitized Belgian newspapers spanning the period 1831-1918.[2] The user inter-

[1] `http://opac.kbr.be/belgicapress.php`

[2] A larger number of pages, subject to copyright laws, are exclusively available within the library

face displays advanced functionalities such as full-text searching, which was made possible thanks to the OCR (optical character recognition) process. Other search parameters include time ranges, specific dates, newspapers titles and languages (French, Dutch or English). The scope of our study will solely focus on BelgicaPress, although it represents only a subset of the main digital catalog "opac.kbr.be". The dataset covers a period of one year, starting from 1st of January 2016 to 1st of January 2017. *work in progress*

The library wants to understand to what extent the portal is being used by the global public and genealogists, who are mainly interested in finding back information about family members or their local communities, or whether the majority of users consist of historians and other researchers who perform queries in relation to specific historical events or well-known personalities. In other words, the library wishes to understand what is the ratio between between a query such as "Ferdinand Foch Passchendaele" (name of a French general combined with the place name of one of the iconic battle fields of World War I), in comparison to queries on names of relatively unknown individuals and localities which never were the scene of any mayor historical event.

From a conceptual perspective, one can not postulate a binary distinction between both types of research needs. For certain queries, the line will be certainly blurry and all depends on the context and the specific need a patron had in mind. However, the authors had to come up with a method to identify a *global tendency* across a corpus of queries which is too voluminous to be analysed manually. Also, once an automated method has been put in place, it can be used at different time intervals. In order to implement this research question, a combination was used of Natural Language Processing (NLP), Regular Expressions (REGEX) and the reconciliation of specific tokens from the queries against knowledge bases and authority files published as Linked Data. For both entities types (person and place names), assumptions in regards to what can be of *global* interest (scientific interest from historians or media scholars) or of *local* interest (information needs from genealogists or the general public) were translated into a set of *rules*.

First of all, we consider a query on a person name to be of local interest if a query contains both a first and a last name, which is not represented in a variety of knowledge bases. The following rules are used to identify the subset:

- Queries containing more than one token
- Queries containing a token which has been matched with a list of Belgian first names, provided by the National Archives
- Queries for which either a two or three gram, consisting of the token(s) preceding or following the first name, can not be mapped to either Wikidata, DBPedia or VIAF

Secondly, we consider queries to reflect an interest in a local location, if it concerns a geographic entity which is not mentioned in more than three Wikipedia pages and if its own page has a wordcount beneath XXX ({**TODO**: Anne and Ettore, please experiment and define the rule accordingly.}

Finalement : évaluer le potentiel des données d'autorité disponibles en ligne pour répondre de façon automatisée à de telles questions. Possibilités et limites...

4 Data

Usage data are collected day after day. In this section, we describe the content and the limits of the three types of usage data that have been collected:

1. the data provided by the Piwik graphical user interface (Piwik Analytics);
2. the raw data collected by Piwik;
3. the log files stored by the database management system.

4.1 Piwik Analytics

Piwik offers an open source alternative to Google Analytics: a graphical user interface, hereafter called "Piwik Analytics". The main difference between the two service providers is that Piwik allows the user to own all the data whereas Google maintains property rights on the data.

Piwik Analytics can be seen as a very practical tool that gives a quick and general overview of the data. Concretely, each time a user searches through a catalogue, the Javascript tag inserted within the webpage is activated and a new "visit" is recorded locally in the Piwik database. The identification of users, which is limited to an ID number, is based on either cookies stored on their computers or their IP addresses. Cookies are small text files that are stored on a user's hard disk and make it possible to identify a connexion between a browser and a server in order to visualise web pages (this is done so that the user's experience can be personalised by).

The raw data stored in the database are aggregated during the archiving process to enable the production of end-user reports. This exempts end-users from processing an enormous amount of data every time a new report is needed (Piwik 2017). The reports, displayed in Piwik Analytics, are classified into four main categories:

- Visitors
  The first category includes a real time visitor log, information about browsers, screen resolution and device used by the visitor, location of the connection and "engagement" (time on site, pageviews per visit, etc.).
- Actions
  This second category refers mostly to top page URLs, outlinks (click on a link to an external website) and amount of downloaded files.
- Referrers
  The third category indicates where the visitor was before visiting a given website. Traffic sources are divided into four types: direct entry; search engines; websites; campaigns (newsletters, for instance).
- Goals
  The last category offers the opportunity to create goals and then to measure the amount of visitors performing a specific action during their visit, such as downloading a file or viewing a digitised file.

Since BelgicaPress is a subdomain of opac.kbr.be, information made available in Piwik reports is unfortunately related to the whole domain and not specifically on BelgicaPress usage.[3] This is one of the reasons why it is useful to look at the raw data.

4.2 Piwik raw data

While Piwik Analytics offers a perfect first glance at the data, the information displayed by the tool is not sufficient for an in-depth analysis: it requires the Piwik raw data. Each time a user browses the website, the visit is recorded by the Piwik server and the raw data is stored locally in a database whose tables can be read through SQL commands. This enormous amount of data, difficult to process at once, can be exploited to answer more complex questions. Moreover, custom variables can be added to address two issues. Firstly, they can be used to complete ambiguous URLs which do not reflect explicitly user actions (Chardonnens, Hungenaert, Vanbrabant 2017), which is not the case here. Secondly, they are helpful to measure specific interactions between users and the interface, such as a full-screen view of a newspaper page in the case of BelgicaPress. Thus, custom variables offer a way to keep track of actions based on Javascript code, which would not have been detected automatically by Piwik.

Data needed for an in-depth analysis is extracted from the Piwik database and stored in a structured file to be explored using computational methods. In the end, six elements have been extracted:

- the visitor ID;
- the visit ID;
- the timestamp: day, hour, minute and second of the action;
- the URL, whose user queries and other information can be parsed ;
- the custom variables containing additional information about the browsing behaviour: full screen mode activation, full view of a newspaper, a click in the list of publications, etc.

These elements are sorted by visit ID and then by time, in order to reshape the visitor path through BelgicaPress pages. Moreover, the data exported may be supplemented by others information such as the browser language or the screen resolution, if deemed necessary.

4.3 Log files

The log files are structured data which are automatically stored by the database management system of BelficaPress. Each time a user browses the BelgicaPress application, a request will be sent to the webserver and a reply will be sent back from the webserver, which has extracted data from the database. That operation is automatically recorded in a transaction journal. Log files can be text files or MySQL records, as in the Royal Library of Belgium context.

[3] Theoretically, it is possible to create a segment consisting exclusively of visits including a BelgicaPress URL, however performances issues were encountered during this attempt, due to the amount of data to process.

To be used for analytical purposes, the logs require major prepocessing steps. First of all, the deletion of web robots visits: contrary to Piwik whose a plug in[4] is responsible to filter web robots (bots, spiders and web crawlers which automatically scan websites), the logs files record information about every visit, be it a human or a robot. Although the use of a bot tracker plugin cannot guarantee completely reliable results, Piwik exempts us from this "cleaning" step. In addition, the logs are not automatically aggregated by visits or by visitors. This major difference can be erased using IP addresses and timestamps to reshape the visits through the website and separate them into sessions (Nouvellet, 2017). However, even if sessions can be recreated, the identification of unique visitor is made difficult in the absence of HTTP cookies: IP addresses are limited when people use a collective access through a proxy server or when they operate in an environment that makes use of dynamic IP addresses (Voorbij, 2010).

In the case of BelgicaPress, two types of logs are available: firstly the data related to a query in the search engine to access digitised newspapers; secondly, the data related to a view of a digitised newspaper page. Here are the more relevant information recorded for each request send to the webserver:

1. User queries
    - the timestamp: day, hour, minute and second of the request;
    - the IP address of the user, with a visible distinction between internal (within the institution) and external IP addresses;
    - the search terms of the query and the potential selected filters (time interval, newspapers titles, etc.).
2. Display of a newspapers page
    - the timestamp: in this case, it contains only the day, neither the precised hour, minute nor second.
    - the IP address of the user, with a visible distinction between internal (within the institution) and external IP addresses;
    - the "uurl", i.e. the unique identifier of a newspaper number;
    - the "idn", i.e. a classfication ID by newspapers and by year;
    - the database, i.e. always the Belgicapress database in this context;
    - the type of file displayed, i.e. always jpg in this context;
    - the page number of the item displayed;

4.4 Selected data source

Considering the possibilities and limits offered by these three types of data, we have chosen to use the Piwik Raw Data to create our dataset. Beyond strictly logistical reasons (the institution had stopped collecting the log files during the year 2016), our choice was guided by four reasons.

First of all, they offer far more precised data than the aggregated data displayed by the Piwik graphic user interface. Secondly, they do not require the pre-processing steps

[4] `https://plugins.piwik.org/BotTracker`

required by the log files to recreate the sessions of each visitor. Thirdly, they present completed data about a visitor session whereas the log files made available to us by the Royal Library contain only data related to user queries or document views. Lastly, by combining IP addresses and HTTP cookies to identify visitors, they provide more accurate visitor numbers than log files. By way of illustration, for a same test period (october 2015), almost 18% of distinct visitors identified by Piwik were not recognised as such in the log files (Log files: 1071 visitors identified via IP addresses, Piwik raw data: 1298 visitors identified via cookies and IP addresses).

Although they prove to be the most exhaustive, the Piwik raw data themselves have limitations. Foremost, usage data, such as user query and the results of that query, is contained in an URL and not presented in a structured way[5]. To handle this and to perform an in-depth analysis, URLs need to be parsed, which means that the relevant data need to be identified within the URL and then extracted in a structured file. On top of that, since the usage data needed for analysis is stored in the URL itself, it may cause complications. For example, the search terms entered by a user appear in the URL. As a matter of fact, they remain present in the next URLs when the user skims the results of his query. It means the presence of the search terms is proportional to the view of results and not striclty linked to a "real" query into the search engine. The next section explains the method we have developed to manage this type of limit.

5 Reconcile user queries with authority data

5.1 Data pre-processing

Data pre-processing included four steps: filtering, parsing, grouping, and cleaning. There are fundamental to reduce the amount of data to be analysed and to obtain data as significative data as possible.

1. Filtering
   In order to prevent performance issues due to large amount of raw data, two strategies have been used. First of all, the survey period has been limited to one year (from 1st of January 2016 to 1st of January 2017). Second of all, we used a text filter during the data extraction[6] to keep only URLs related to BelgicaPress and thus exclude others URLs related to the main catalog of the Royal Library of Belgium. This led us to a text file of 234,6 MB, containing 1,099,043 lines. Here are examples of the three types of URLs extracted:

   Homepage `opac.kbr.be/belgicapress.php?lang=NL`

   Results page `opac.kbr.be/pressshow.php?adv=1&all_q=&any_q=&exact_q=LondonDiamond&none_q=&from_d=&to_d=&per_lang=&per=&lang=NL`

   Pageviews `opac.kbr.be/pageview.php?all_q=&any_q=&exact_q=LondonDiamond&none_q=&from_d=&to_d=&per_lang=&per=&sig=JB555&lang=NL`

[5] It has to be noted that Piwik offers functionalities to track internal search keywords and obtain them more easily. These functionalities have not been implemented in the context of our project, but could potentially facilitate the process explained in this paper.

[6] A regex inserted within the SQL LIKE operator.

2. Parsing
   The second step consisted of parsing the URLs to extract the search terms. It has to be noted that the Belgica Press interface allows the user to specify the type of his search terms. He can combine them in different manners: he has the possibility to search "all of these words"; "one of these words"; the "exact phrase" and can also potentially refine them with a "none of these words" query.

   The task has been performed by firstly reading the text file and convert it to a dataframe, using Pandas, the Python data analysis library. The relevant data have been extracted in a new column with the help of regular expressions and concatenated when several queries have been combined. During that stage, a minority of queries (less than 2% of the total) were wrongly lost, in particular because of the presence of "&" in the queries themselves, the special character which is used as delimiter to parse the URL.

3. Grouping
   As aforementioned, queried terms entered only once during a visit will appear several times within the next URLs if the visitor consults the results pages. Counting all these occurrences indifferently would therefore skew results such as the most frequent terms.

   An approximate but nevertheless consistent method consisted of keeping each distinct query only one time per visit. The pandas "group by" operation has been used to group rows containing the same querie(s), based on the ID of the visit, the queried term(s) and the timestamp. The timestamp has been used here as an HTTP request identifier: it eases the distinction of several occurrences of the same query (strings are identical) during the same visit (the ID numbers are identical).

4. Cleaning
   Before being analysed, user queries need to be "cleaned" in order to enable us to associate similar string of characters despite superficial differences (Manning, Raghavan & Schütze, 2008). To unmask the "hidden duplicates", several actions have been taken: to trim leading and trailing whitespace; to collapse consecutive whitespace; to convert all characters to lowercase; to replace each character which is not alphanumeric by a space; to replace special characters (for example, replacing "à" by "a"). The lack of context raising already ambiguity issues, the decision was made to not chop queries into separate entities ("tokenisation"). By contrast, queries containing less than 3 characters, have been judged not meaningful and have therefore been removed. The grouping operation was processed once again after this cleaning step.

At the end of these data-processing, the dataset contains a total amount of 83 854 queries, which is composed of 52 547 distinct queries. A little less than 30 000 visits on Belgica-Press website (29 812) are at the origin of these 83 854 queries: this leads us to an average of 2.87 distinct queries per visit, with a standard deviation of 4.1 and a median of 1. The minimum is 1 and the maximum is 107 queries/visit. The number of tokens per query ranges from 1 to 64, with an average of 1.8, a standard deviation of 1.08 and a median of 2. Moreover, it has to be noted that 98 percent of the dataset (82 279) contains 5 tokens at most.

5.2 Extracting geographic references

In this subsection we describe our pipeline to automatically extract location names and estimate their "degree of locality". To extract geographic references, the first step consists of tokenisation. Once the query subdivised in tokens, the key is to match each token or as many tokens as possible with one of the location names contained in a authority file. Thus, that step allows us to match not only "Brussels" with "Brussels", but also longer queries such as "Mont Sainte Aldegonde" with "Mont-Sainte-Aldegonde", which was written without hypens in the original query and is therefore separated in different tokens.

The authority file we used in this context is a dump from the geographical database GeoNames covering all Belgian places in a broad sense. A preliminary analysis having shown that location names are mainly Belgian, decision was made to focus only on Belgian names, to optimise processing time and to reduce noise due to irrelevant results.

The flat file contains approximately 30 000 different names, including different spellings as well as different language versions. It means that for a single place like Hoeilart, more than ... alternative versions will also be recognised, increasing significatively the likelihood of reconciliation.

After reconciliation, the data remain easily accessible and manipulatable, with the additional possibility of having the exact coordinates for a given location, as well as a hyperlink to the interactive map of GeoNames.

In spite of the general success of the method, some ambiguity and misspelling issues need to be underlined. Firstly, several municipalities possess the same name. Thus, a query containing the place name "Saint-Nicolas" could be associated with different Belgian municipalities called "Saint-Nicolas" or even "Sint-Niklaas". There is no way of knowing which place was designated, except to dive into the end user's mind. No to mention a more general problem of ambiguity: did the user was looking for a village called Saint-Nicolas or festivities and traditions of December 6, related to Saint Nicolas.

Secondly, geographic references have to be written in a strictly identical manner in order to be recognised. Thus, "Anwterp", the misspelled name of a Flemish town, will not be matched with the correct spelling "Antwerp". Fuzzy matching algorithms can help to deal with that kind of issues, by enabling approximate string matching. Some tests have been done on our dataset, leading logically to a higher number of matches. However, we have renounce to use one of these algorithm: beyond the fact that they significantly slown down the calculation speed, they tend to provoke an excess of false recognitions.

Once the geographical references contained within the queries have been recognised by using the GeoNames database, they have to pass the "degree of locality" test: do they are mentioned in the label of a Wikipedia page in French, Dutch or English, and, if that is the case, does the page contain more than .... words? That step, simple in appearance, proved to be more complex, due to the combined effect of the natural ambiguity of place names and the inconsistency and incompleteness in the Wikipedia/Wikidata classification itself.... EXAMPLE : Houx

5.3 Extracting person names

The pipeline to automatically extract person names and evaluate their "degree of locality" is similar but less demanding. The first step to process queries and identify those containing a full name of person (first name and last name) consists of isolating those

composed of more than one token. All potential full names will be in the resulting subset. After matching those queries with a (cleaned) file provided by the State Archives of Belgium and containing almost 70 000 first names, we are able to extract first names as well as the preceding or following token, i.e. the supposed surname, e.g. HUBERT Collin. Attention has to be paid to include names involving a particle (e.g. van, von, de, van den, van der), which are composed by more than 2 tokens. The second step implies to map all the potential full names with Wikidata, DBPedia and Viaf, three free knowledge databases. According to our rule, if their names are present in of these databases, they can be considered as "known" persons. The mapping is done using the database API... (à mettre à jour) The third step aims to identify among remaining entities those which can be considered as "unknown people" (e.g. Adeline Pollet) and those which are false positive (e.g. Dieudonné cyclisme). The distinction is made by considering as false positive those whose supposed last same is both present in a lexicon of proper names and absent in a Belgian family names file provided by the State Archives of Belgium. The remainders are considered by defaut as "unkwnon people".

When extracting person names, the main difficulty is related to lexical ambiguity. Indeed, first names can also be common names, such as "Fleur", which can be a proper name or the French translation for "flower". A frequently cited case is "Paris": to be able to extract "Paris Hilton" from a query, the file containing first names should include "Paris". However, this inclusion would imply "noise", i.e. a lot of queries about the French capital instead of a person. This is the reason why the file was first cleaned, even if it means losing some "false negative". In addition to this, specific cases arise like the name of athletes, military or historical figures, which do not consist of the traditional structure "first name" + "last name": Constant le Boucher, Commandant Soleil, Colonel van Boogaert, Leopold III or Comte de Chiny.

Furthermore, when evaluating their degree of locality, we test if the complete name is present in the French, Dutch or English Wikipedia pages. Yet, the content present on these pages is far from exhaustive and will inevitably bias the results.

6 Evaluat

6.1 Creation of Gold Standard Corpus

To evaluate the different steps of the pipeline of NE recognition, a manual analysis was required. The aim was to examine which proportion of queries actually contains named entities, by annotating a representative sample of 1 000 randomly selected queries. Considering the two specific research questions related to place and location names, it was decided to focus only on these two types, although other types may be retrieved, such as organisations.

These two categories - persons and locations - have been extended to face ambiguity issues, resulting in 5 categories:

- LOC for locations in a broad sense: any geographical location corresponding to a place, be it a municipality, a country name or even a subway station (e.g. "Horta station").
- PER for a full name, a last name or just a first name (e.g. "Leopold II"). When relevant, the presence of a full name (first name and last name) has been was reported in an additional column.

- PER LOC for ambiguous cases where the entity may designate both a place or a person (e.g. "général Jacques", which turns out to be at the same time the name of a Belgian soldier and a street in Brussels).
- PER AMBIG for entities which vaguely look like a person's name, but no known place or person can be associated with (e.g. "Tombek").
- AMBIG for very ambiguous tokens: it could be an entity named PER or LOC as well as another type or a common name (e.g. "stampe", "valk").

Moreover, a general rule has been set: no overlap. Thus, "August van Turnhout", which is clearly a full name, will be annotated as such, while "Turnhout" will not be annotated separately as LOC, although it is a Belgian locality.

The annotation task intended to dive into the user's mind to try to understand what he was looking for. It means that the trained annotators performed the annotation using contextual data (the other search terms entered during the same visit), to obtain information as accurate as possible and reduce ambiguity. Thus, considered alone, the query "Corbiere", is quite vague. A glance at the previous and subsequent queries ("de la Corbiere" and "Lacorbiere") let us assume that the query probably refers to a PER entity (the french painter "Roger de la Corbière") and not "Corbières", the Swiss location or the French wine of the same name.

Once emptied of duplicates and unexploitable queries (queries composed only by numbers), the sample consists of 995 queries corresponding to 952 visits. The 995 queries were annotated by two of the authors, with the help of online search engines and databases such as Wikidata or Geonames. At the end of the process, divisions of opinion were discussed with a view to reaching a consensus. When no consensus was reached, each author retained his initial annotation: one annotated the "Lambeaux Horta" query as a place ("Le pavillon Horta-Lambeaux", a building in the Parc du Cinquantenaire in Brussels), while the other saw juxtaposed patronyms of the Belgian sculptor and architect.

6.2 Cohen's kappa

	Annot.1	Annot.2	Consensus
PER	485	477	473
LOC	317	317	313
PER AMBIG	18	25	17
PER LOC	18	19	16
AMBIG	10	11	10

Sur 849 entités, 829 (97.6%) ont donc été classées dans la même catégorie par les deux annotateurs. Malgré quelques dissonances, la méthode de consensus utilisée aboutit donc à un accord inter-annotateurs très important, avec un kappa de Cohen supérieur à 0.96 sur un maximum de 1.

Parmi les 313 lieux LOC issus du consensus, 225 (78%) sont situés en Belgique, loin devant le République démocratique du Congo (9%), la France (6%), les Pays-Bas (5%) et une quinzaine d'autre pays comptant chacun moins de dix occurences.

A ce stade, il est intéressant de noter que les 225 lieux belges se répartissent à 86% en noms de communes et sections de communes (presque moitié-moitié : 99 sections et

96 communes sur 225), le reste étant des "Points of Interest" tels que "église Saint-Paul" (10.6%), des noms de sous-régions et de provinces, de forêts et de rivières, etc.

En nous concentrant sur les noms de communes/"villages" belges, nous couvrons donc 62% du total des noms de lieux cités dans l'échantillon de requêtes, tous pays confondus. Sur l'ensemble du corpus, nous devrions donc avoir 95% de chances de couvrir entre 56 et 68% de l'ensemble des lieux cités.

6.3 Analysis of precision and recall

1. places

   S'agissant des noms de lieux belges, le second script Python en a reconnu correctement 183 sur 225 (R=, P=), mais 180 sur 195 s'agissant des noms de communes/sections (R=, P=). Dans ce dernier cas, les principales omissions sont dues à des abbréviations ("wez" plutôt que "Wez-Velvain") ou à une orthographe approximative ("atwerpen" pour "Antwerpen"), voire à peine reconnaissable ("overyss-chf" pour "Overijse"). Afin d'éviter un excès de fausses reconnaissances, ainsi que pour limiter le temps de calcul, nous avons en effet renoncé à introduire dans les deux scripts un algorithme de fuzzy matching, préférant nous fier à la longue liste de graphies alternatives qu'offrent les listes (gazeeters) sur lesquelles les scripts se basent.

2. personnes

   Parmi les 516 noms de personnes issus de l'annotation consensuelle (PER, PER LOC et PER AMBIG confondus), nous en avons annotés 141 (27.3

   Le script Python a identifié quant à lui 178 "noms complets", dont 131 sur les 141 ainsi annotés (les 10 manquants ont été plus ou moins extraits, mais de manière erronée, comme "van loo prosper" devenu "loo prosper". Un autre type problème survient lorsque une seule requête juxtapose deux noms de personnes, par exemple "Victor Hugo Albert Einstein").

   Nous avons donc un rappel (R) de 92.9

7 Discussion

8 Conclusions and future work

The available literature regarding the use of web analytics tends to be quite categorical in either supporting or downplaying the phenomenon and neglects to enter in a detailed discussion of the complexity of large volumes of usage data. We have tried to fill that gap by presenting both critical and supportive elements, backed up by a concrete case study. Furthermore, we have bypassed a positivist theory verification, which measure the information retrieval effectiveness of a catalog in a quantitative manner, by reflecting on the inherent complexity of aggregating and interpreting usage data. -piste pour enrichir les métadonnées -aller plus loin avec Google Books (ex. Augustin de Ley, rapport judiciaire)

9 Acknowledgements

The authors would like to extend their gratitude to the Royal Library of Belgium. They are particularly grateful for the assistance given by Erwin Van Wesemael and Nicolas Roland. The support and helpful feedback of the promotors of the ADOCHS project have also been invaluable to the success of the research and the conception of this article. The authors would therefore like to thank Ann Dooms, Florence Gillet and Frederic Lemmers. The research underlying the results presented in this article was funded by the Belgian Science Policy Office in the context of contract number BR/154/A6/ADOCHS.

References