The unfathomable secret remains: a reverse engineering study of SEO factors on Google's Search engine

Working Paper

2014

Antonio Cañabate Carmona, Department of Management, UPC · BarcelonaTech.

Pau Fonseca i Casas, Statistics and Operations Research Department, UPC · BarcelonaTech

Ferran Sabaté Garriga, Department of Management, UPC · BarcelonaTech.

Abstract

The vast majority of users don't seek results beyond the second page offered by the search engine, so if a site fails to be among the top 20 (second page), it says that this page does not have good SEO and, therefore, is not visible to the user. The overall objective of this project is to conduct a study to discover the factors that determine (or not) the positioning of websites in a search engine.

Keywords: SEO, Search Engines, PCA analysis, Cluster analysis

1. Introduction

The Internet has been consolidated as a distribution channel along last decade; consequently internet marketing has become a key factor for firm performance (Geyskens, Gielens, & Dekimpe, 2002). Both, e-commerce and internet marketing have experienced a strong growth since historical annual revenues grew from 6 billion dollars in 2002 to 31.7 in 2011 (Silverman, 2012). For internet marketing activities, it is observed searches and display-banners are the preferred format for ads (47% and 22% respectively). In (Silverman, 2012) searchers includes different marketing strategies categories (Paid listings, Site optimization, Contextual search and Paid inclusion strategies) related with the word (or words) used by the user for searching or with the content of a web page.

Firms involved in e-commerce have to work hard in being visible and increasing its website traffic in such enormous market that the Internet is. They have to understand how users search content, products or services through search engines (i.e. Google) or infomediaries (i.e. shop bots). (Papatla & Liu, 2009a) findings give support to search engines and infomediaries as group of actors who are playing a strong role for leading consumers to e-commerce web sites. They also conclude the role of search engines could be twice important than those played by the infomediaries. About keyword relevancy, (Suchanek, 2010) study points out common keywords are well related with e-commerce.

Firms website must to be listed as top as possible in search engine results pages (SERP), or at least, within the first page. To achieve this objective firms must include in their internet marketing strategy, a set of techniques focused on improving their website position. These techniques used to be classified as Search Engine Marketing (SEM). (Paraskevas, Katsogridakis, Law, & Buhalis, 2011a) proposed a global framework for SEM strategy development, made up of four stages: analysis, planning, implementation and control. (Sen, 2005) referred four actions within SEM strategy: keyword-related banner advertisements, paid submission/paid listing and paid inclusion for regular update, search engine optimization (SEO), and paid placements (PP). All previous techniques except SEO imply a direct payment to the search engine company. On the other hand, SEO strategy doesn't require a direct payment, although its implementation cost depends on how much time personnel of firm works to improve its website code, content quality and the network of in-links from other pertinent websites. As our study is focused only in SEO, readers interested with paid techniques can consult some useful references (i.e. (Pan, Xiang, Law, & Fesenmaier, 2010; Paraskevas, Katsogridakis, Law, & Buhalis, 2011b; Sabaté, Cañabate, Velarde-Iturralde, & Griñón-Barceló, 2010; Sen, 2005; Sullivan, 2002).

Search Engine Optimization is a set of procedures and techniques to study the characteristics of a webpage to reach maximum visibility through search engines. SEO allows firms to increase its website traffic by increasing the number of visitors as it has been empirically proved (Lee, Kang, & Chung, 2012). According to (Evans, 2007) there are mainly two kinds of factors influencing the website position within SERPs: Query-Factors and

Query-Independent. The first one is related to both how the page is internally built and its content. The idea is to make easy for the search engines robots the understanding of what is told in the webpage and what keywords fit better. Thus, it is important to optimize the webpage using appropriate keywords to make a link between its content and those keywords more probably used by users when they search through search engines. Xue et.al (2009) offers an overview of how to develop a website according to main in-site factors. On the other hand, Query-Independent factors are related to information from those other web pages which link to ours, in order to maximize our impact factor, trust and reputation factors, among many others. A better value in such factors a better position within SERP (Google 2013a).

Given a specific search engine, it would be absolutely useful to know how its algorithm works for positioning a webpage within SERPs. What people know is that many factors influence the place assigned to a webpage. They also know algorithms are dynamics, that is, they are changing constantly and increasing their complexity (Killoran, 2013; Lam, 2001; SEOmoz, n.d.).

Taking in consideration the described context, where the search engines play a key role as intermediaries and firms are increasing the utilization of SEM and SEO strategies, we believe it makes sense to conduct this study which is focused on achieving a better understanding of how Google search engine algorithm works and what factors are relevant. Google algorithm is chosen due to two reasons: (1) Google is the most used search engine, which has reached a 90.6 percent of use (being Bing the second one with 3.49 percent) between January and April 2013, according to (StatCounter, 2013); (2) the study is based on data from SEOmoz which was collected through Google search engine after making 12,000. Next, in section 1 a literature review can be read for a better understanding of SEO context. An exhaustive description of the analyzed data is described in section 2. Principal component and cluster analysis and results are described in sections 3 and 4 respectively. Next a discussion is exposed in section 5.

1. SEO context

On the Internet, searching through search engines is the most important activity according to users (Papatla & Liu, 2009b). Internet users find content describing what are looking for by mean of keywords. A keyword can be a single word as "car", multiple words as “red car” or a phrase as "buy cheap cars". Firms know internet users work in this way when they browse the World Wide Web. Therefore, after establishing a set of keywords that fit as well as possible with their product or service, firms have to design their website using those keywords in two ways: within content and in technical items of HTML code. Only in this way it will possible to be listed in first positions within SERP when somebody searches using those keywords. In fact, keyword is the term for which it pursues the optimization of a website.

The way users interact with search engines is well studied. Yet in 2001, (B J Jansen & Pooch, 2001) reviewed what was published about this research line by that time. They have also proposed a framework for future review. From then, some issues had been studied. (Lorigo et al., 2006) drew a picture from a sample of college students, where users did few queries and put attention to the first ones, although sometimes in a different order from the displayed. According to (B J Jansen & Spink, 2006) users that only processed results from first results pages was over 70%. It fits well with (Bar-Ilan, Keenoy, Levene, & Yaari, 2009) conclusions who considered as most important quality factor the placement within the SERPs, more than the content globally displayed. (Hariri, 2011) considers that users of the search engines must examine at least three or four pages in order to improve their options of finding what they are looking for. In any case, results from optify (2012) put on the table how important is to be listed in the first positions within SERPs. Their results show that the Click Thru Rate percentage along the ten top rank positions decreases in this way: 36.4%, 12.5%, 9.5%, 7.9%, 6.1%, 4.1%, 3.8%, 3.5%, 3.0% and 2.2%. First position probability to be clicked trebles the second one.

Also is observed on (B J Jansen & Spink, 2006) significant differences about using Booleans operators between US or European users, being the Booleans more used by US ones. (Lorigo et al., 2006) also founded that query evaluation was influenced by gender, as for instance, male went further reviewing the listed items while females did more regressions to previous items. And (Bernard J Jansen, Brown, & Resnick, 2007) founded a preference for clicking on non-sponsored results before than sponsored ones, being proved a positive interdependence between paid and organic results too.

From search engines firms point of view it is important to know the best way for making short summaries. (Liang, Devlin, & Tait, 2006) designed a metric for measuring summaries quality. It was based on the representativeness and judge ability according to searchers point of view. Their system obtained better results that those from Google SERPs. Literature findings are not conclusive. (Bar-Ilan et al., 2009) concludes a preference for original order of search engines than other specifically proposed. It could mean that search engines are working on improving their summaries quality. In fact, it is demonstrated brand plays a key role in how consumers search (Bernard J Jansen, Zhang, & Schultz, 2009). Users trust the most recognized search engines although they seem more involved in the search process when they work with less-popular ones.

1.1. SEO Factors

Given a query, search engines make an index of results before printing the results pages, according to a set of parameters or factors. These factors are known as SEO factors (Ref). As said by Google itself their algorithm includes more than 200 factors (Google, 2013a).

The understanding of what SEO factors are used by a search engine algorithm is crucial form firms to be ranked in first positions. Therefore to achieve the top position by a web page, we

might think that we should just focus on improving these positioning factors. However this practice is not simple as the only ones who know exactly which are these factors, and how are these factors used, are their creators. And search engine employees keep in strong secret that information, although sometimes they reveal some tips on how to build websites and interesting content, probably according to their own interests (i.e. Cutts, 2012; Singhal , 2011). In addition it is believed that some of the factors are not under control of websites' owners or even are specific for each particular search (Refs).

SEO factors are determined by the specific version of search engines algorithm. Google changes its search algorithm up to 500 times each year according to SEOmoz (2013a). For major updates a codename is given. By instance, last Google major updates have been Panda (February 2011) and Penguin (May 2012).

Given that search engines algorithms are secret, the academia and those experts from Search Engine Optimization industry are working hard to find out which those SEO factors are. Commonly authors classify the factors in two categories: query-factors, for those related to the web-page content, and query-independent factors, for those related with data from the in-link (i.e. Bifet et al., 2005; Moran et al., 2006). Alternative terminology refers to the same concept as "onsite/on page" or "offsite/off page" factors (i.e. Rimbach et. al., 2007), among others. SEOmoz (2011) proposed a richer classification based on nine categories. Four are defined at domain level (Domain Level Keyword Usage, Domain Level Link Authority Features, Domain Level Keyword Agnostics and Domain Level Brand Metrics) and five more at page level (Page Level Social Metrics, Page Level Link Metrics, Page Level Keyword Usage, Page Level Traffic Data and Page Level Keyword Agnostic). Our analysis will be explained according to these categories.

From academic studies we have achieved partial information about which SEO factors could be. Khaki-Sedigh and Roudaki (2003) made a prediction model for the Google ranking dynamics based on a linear regression. They included these factors: PageRank, Keyword frequency in the web pages visible text, Keyword density in the web pages title, Keyword density in the web pages text, Keyword density in the web pages linked text, Keyword density in the web pages ALT tags, Keyword prominence in the web pages text and similar pages. They did not justify which one could be more important.

Bifet et al. (2005) probed some ranking functions for comparing their results with the Google ones in order to achieve a better understanding of the Google ranking algorithm. They study took in consideration 22 factors classified in four groups (content, formatting, link and metadata). It is remarkable that their design for queries was based on four categories, which fits well with the idea that the Google's algorithm could work different depending on categories of keywords or topics. These categories could be related with topics or "types of questions" that the search engine could process in a different way. From the factors considered, the "Number of pages linking to a page" was the most important for two categories and the third in importance for another one. "The page rank" was the

second in importance in three categories. “Similarity of the term to the document” was the most important for that category related with the use of 6 keywords simultaneously on the domain of statistical inference. And “Fraction of terms in the documents which cannot be found in an English dictionary” was the most important in one category and the second one in another, being its correlation negative.

Fornutato et al. (2005) argue that the global nature of PageRank make difficult predict the position where a web page will be ranked. Despite of that assumption their study concluded the existence of strong correlation between both factors “Number of pages linking to a page“ and PageRank.

Evans (2007) analyses the effect of some factors in Google’s algorithm. The study considers a short list of factors (PageRank, number of pages in a site indexed, number of in-links, domain age, number of domain links, number of pages listed in Del.icio.us, the Alexa traffic rank, and to be listed in Dmoz directory. Results could show a positive effect of having a higher PageRank, number of in-links, and number of bookmarks in “del.icio.us”. To be listed in Dmoz directory and an older domain age seem to point out to a better ranking. The study also explains that SERP rankings depend on location since Google data centers around the world are not always fully synchronized. For that reason the same query could give different results if it is launched from different countries at the same time.

Previous references make clear the key importance of the PageRank factor in Google’s algorithm, already from year 2005. Rimbach et. al (2007) argue that a better understanding of how the PageRank is calculated and implemented is necessary in order to allow marketers to improve their SEO strategists. PageRank calculation and its implications for SEO marketers are considered in two scenarios: the original one, where the relative “importance” of web-pages is calculated with independence of the search query; and the topic-sensitive scenario, where the algorithm takes in consideration those topics that fit better with those keywords in search query. The topic sensitive rank factor is based on Haveliwala (2002) previous work. In the topic-sensitive scenario a good SEO strategy for a company would be to adjusting its links exchange programs with the purpose of having in-links coming from a distributed network of web sites, being their topics compatibles with the company ones.

The scheme for calculating a sensitive-topic PageRank could be relevant for this study. If the same scheme is applied to calculate or interpreted other factors, it could point out to a different relevance of SEO factors depending on topics.

PageRank among other criteria allow Google to determine the popularity of a website. Thus, higher popularity higher ranking. Atrey et al. (2012) argue that PageRank could not be a good factor for considering the level of trust, since could there be a plenty of credible websites with low popularity level, which could be not listed in top positions in SERPs. The authors conceptualize trust more based on factual aspects of information than on private or

security issues. Therefore, their contribution was a computational model to determine the trust based on semantic similarity with other websites, considering text and multimedia elements. Their method share inherent features with the one proposed by Bizer et (2004).

It is unquestionable than search engines must be consider some factors related with trust. Since around 2005 experts have been talking about the TrustRank (i.e. Gyöngyi 2004; Wall, 2005). Their calculation is based on a set of seed pages labeled as reputable by any expert and the structure of in-links to a specific website. Higher number on in-links coming from valuable pages o pages itself pointed out by the valuable ones, implies higher TrustRank. Even SEOmoz computes the Moztrust (SEOmoz 2013b) inspired in TrustRank. Despite of these speculations, Google itself talk about trust (Cuts, 2011). In sum it seems like trust, reputation an authority are related concepts that Google try to evaluate through different factors and algorithms and PageRank is one of them. A partial transcription of his speech is quoted here to illustrate that argue:

> "… So PageRank is the most well known type of trust. It's looking at links and how important those links are. So if you have a lot of very high quality links, then you tend to earn a lot of trust with Google….
>
> … But you can kind of break them down into this notion of sort of trust and how well you match a particular query. So how topical you are……Just on the merits of what the user typed in…
>
> ... And it's not that we have something specifically called trust rank, or we have something specifically called authority rank, or something like that. We're basically just trying to say, in the general scheme of things, how much reputation, or how much are we willing to believe that this is a high quality page, or a high quality site?...
>
> …So we use a lot of different words like trust, reputation, authority. And PageRank is a specific example of those sorts of things…"

Web traffic is another factor that has been studied (Kowalski, 2010). A higher user activity on website a higher ranking is expected. What make difficult to test this conjecture is that search engines don't like automatic queries from the same IPs since it is understood as abuse. Therefore authors design a multi-agent system to increase traffic in a website imitating those parameters that fit well with human activity. It is discussed how a method like this could improve the website traffic and its ranking, if one assumes the risk to be penalized.

Killoran (2010) analyses the search engine marketing conducted by small firms throughout their technical communication services. He studies the webpage title as SEO factor, concluding that the inclusion of the name of the company and words such as *welcome* or *home* are not useful for improving the website location in rankings. Like Bifet et al. (2005),

fortunate et al. (2005), Evans (2007) and Kowalski (2010) he observed the positive effect of increasing the in-links to the own website.

Further on SEO Factors, search engines seem to prioritize owner services before the competitors' ones in their rankings (Hochstotter and Lewandowski (2009). As evidence of this conclusion the authors compared how many times Youtube links were included in top positions of SERPs after querying through many search engines. Google included Youtube links about four times more than Yahoo and five times more than MSN. The web browser itself seems to be a relevant factor for rankings, because Google take into consideration the sociological patterns of the users of specific browser (Killoran 2013). Thus, the same query at the same time launched from different browsers could offer a different SERP.

Knowing the importance of factors as website traffic, in-links and PageRank among others, SEO marketers are tented to implemented black hat techniques in order to increase their website in rankings. SEO black hat techniques are those that visibly violate the guidelines of conduct published by the search engines (Google 2013b). Malaga (2008) examines some of these techniques, for instance, Blog-Ping, Cloaking, Doorway pages, etc. Blog-Ping is based on making lots of new blogs with links to the own website and then updating its content often in artificial way. Cloaking consists in designing two versions of website, one specific-oriented for crawlers and another for humans. Doorway pages technique consist in designing a set of pages where each one is dedicated to one keyword and has links to the owner website. All these techniques have in common the idea to be excellent in specific factors considered by the algorithm by means of artificial processes. The problem is that search engines consider their utilization as an abuse since these techniques pretend to confuse their algorithm ability to identify which websites are trustier and relevant. Therefore penalties are applied to those firms which Google identify applying such techniques (Cuts, 2013) as for example that one fined to Interflora (Goodwin, 2013) for placing about 150 advertorials on regional news sites.

To conclude this revision of academicals contributions focused on the study of SEO factors, it fits to cite an exhaustive guide of SEO Techniques for web developers (Killoran, 2013). The guide emphasizes three issues: identify keywords considering the target audience and competitors; including the keywords in those elements that are showed in SEPS by Search Engines; and make your website a relevant node in the network of websites related with your topic, in other words, involve other content creators. Going deep in these issues, the guide justifies and explains how to implement specific SEO techniques.

1.2. SEOmoz

Search engines do not reveal information about their algorithms for many obvious reasons, such as the apparition of new competitors or avoiding Black Hat Search Engine Optimization (BHS, is defined as the set of techniques that are used in order to achieve higher search rankings in a fraudulent manner). In fact, search engines try to hide this information. For

these reasons search engines reveal only tips, according to their interests, on how to build websites and interesting content for the user.

Because throughout history there have been no cases of leaks, the only information that is available today regarding search engine optimization algorithms are the opinions of the experts. These experts base their knowledge on the experience gained over the years, so that those views do not always converge.

One of the most reliable sources of information regarding SEO we can find is SEOmoz.

Founded in 2004 from a traditional marketing company, SEOmoz is currently the leader in the field of SEO in Internet. Its main activity is to provide consulting services SEO (Search Engine Optimization) through the development of several applications. Its most important application is SEOmoz Pro, which allows the user to obtain improvements in web positioning in a simple way apparently.

One of the most important characteristics of SEOmoz and brings us most for this project is the publication in 2011 of a study concerning the search engine optimization and positioning factors influencing it. Mention that SEOmoz published a similar study in 2009, but for the realization of this project will be used the most current data, therefore from now on we will refer to the study of 2011.

SEOmoz analyzed and coordinated the opinion of 132 professionals in the SEO world by conducting a questionnaire. Developed a series of closed-ended questions, (3 to 4), each of the experts should answer on an individual basis based on their experience and personal judgment.

With the results of these surveys SEOmoz produced a report showing which are the most important SEOF when positioning a page (SEOMoz, 2011). Additionally factors were grouped into categories and they calculated the influence of each category.

However, SEOmoz do not publicity the format used for the survey, nor the weighting system for each question. After all, these two parameters determine the numerical value of the factors of the report, so the study is subject to the criteria of SEOmoz.

Additionally in exceptional way SEOmoz publics besides the traditional study of the opinions of the experts, a collection of measurements for the SEO factors respect to a set of searches done recently for a specific web. These data are analyzed by SEOmoz using Pearson correlations relating the position of the search for a particular page. This project aims to explore the possibilities of further analysis of these data with different techniques.

2. The data

About 800 keywords for each of the 15 categories are initially selected for the study. In this way are obtained 12,000 different queries to study (800 x 15 = 12,000). From these 12,000

queries, the first step is to filter those keywords that similarity may be present in two distinct categories, i.e., repetitions of keywords present in more than one category are deleted. After this first filter, we get approximately 11,000 different queries. From Google.com are selected the first 30 results obtained, ordered according to their classification. Are discarded all those results that are not of type url, i.e., all those results with formats: .doc, .pdf, .jpg, etc... In addition, if a query gets less than 15 results is eliminated of the sample.

Once obtained the results of all queries, and applied the previously commented filters, we obtain a total of 273,000 different urls. Figure 1 represents graphically this process.

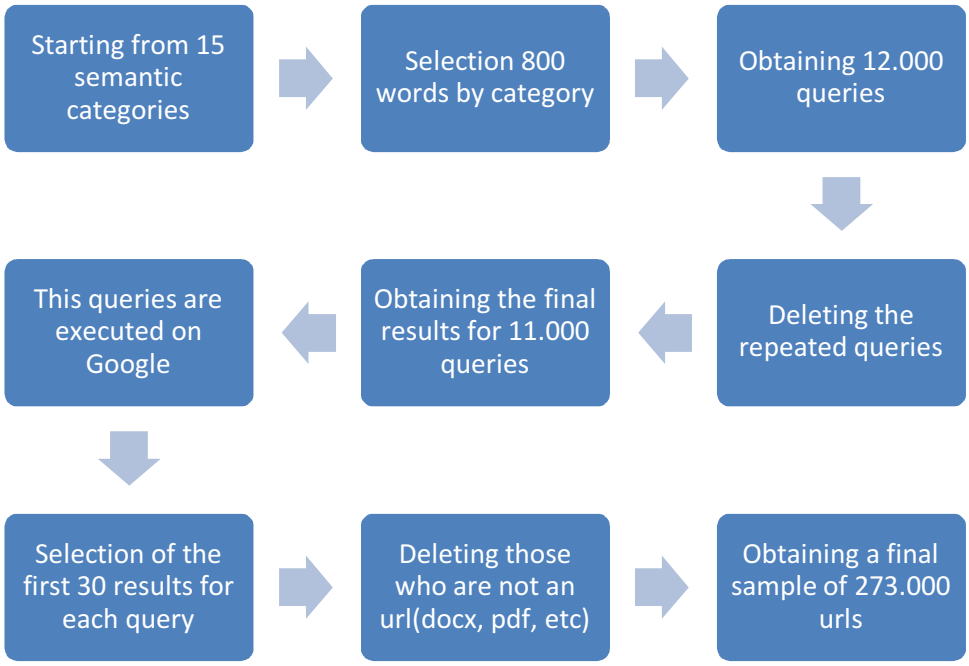


Figure 1. Data acquisition process followed by SEOmoz.

The study is later than March 2011, with which data are obtained after the last major update of Google. All the keywords are in English to ensure the independence of the results; the used keywords are separated into 15 different categories. This decision ensures that the results obtained are general results and are not, by contrast, focused on an area or category. All the keywords that make up each of these categories have been selected with the help of *GoogleAddwords*, the tool that uses Google for advertising. The categories are (i) outfits for clothing (clothing, accessories, etc...), (ii) beauty and personal care (iii) computers (iv) consumer electronics in general (v) finance (vi) food (vii) gifts and occasions (viii) greetings (xix) leisure and free time (xx) home and garden (xxi) events Media (xxii) real estate (xxiii) sports and fitness (xiv) travel and tourism and (xxv) vehicles.

Although a priori the most natural idea would be choosing the most commonly used keywords, i.e. the keywords most commonly used when searching on Google for each

sector, this strategy has not been followed. Is attempted to select a set of keywords with participation in all volumes of searches, i.e. have been selected both popular keywords frequently every month, such as keywords that are not used in a way so regular. Thus continues the idea of global results. Specifically the selection of the keywords that we finally use for the study follows the pattern of **Error! Reference source not found.**.

Table 1. Query selection depending on the keywords popularity.

Number of queries by month	Number of keywords in the study
<1.000	723
1.000 - 5.000	3.574
5.000 - 10.000	2.875
10.000 - 20.000	1.435
>20.000	1.664

Once studied and analyzed the data set obtained by the latest survey from SEOmoz, as well as their origin, we can expose the following conclusions:

1. The detail of the meaning of each of the 164 measured factors is available. Not about its calculus, but its description, classification, and final value.
2. The published data represents a significant sample to perform a further analysis.
3. The data collected are not a subset of searches determined on an area or subject; they are significant on all searches performed in a search engine. As mentioned above, they are exhaustive data semantically and that contains a variety of positions (first 30 results).

Based on the obtained conclusions the matrix of data we are going to use in our analysis is composed by 167 columns and 273.000 rows.

2.1. Missing data

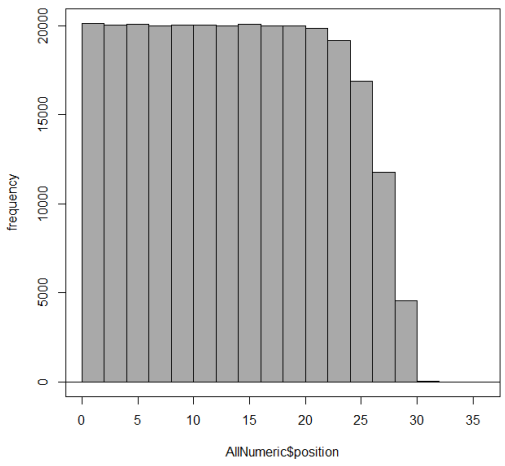
The acquisition of the information is subject to different issues and technical limitations. As an example if during the process of capture a new factor from a website, this website is down, it could be impossible to obtain this information. This lead us to the need of analyze the missing values that appears on the data. First we need to understand the shape of the distribution of the missing values regarding the interest variable (position), see Table 2.

Dealing with missing data led us to analyze if we are in Missing at Random (MAR) Missing Completely at Random (MCAR) (Heitjan & Basu, 1996) or dependent scenario. When data are MCAR, there is no relationship between the probability of the missing data on a given variable and any other variables in the dataset, in that case we can apply *listwise deletion*.

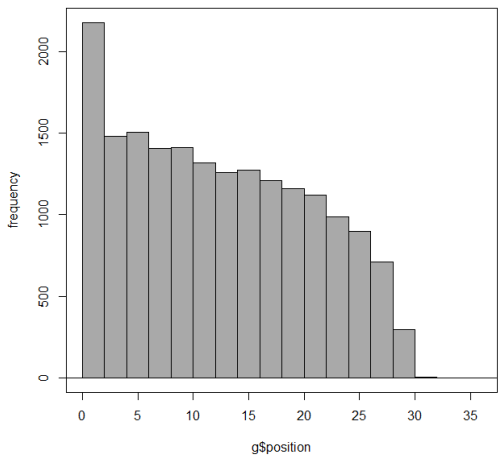
Table 2. Missing values regarding the observations.

Position	1	2	3	4	5	6	7	8	9	10
Percentage of urls with some missing	89%	93%	94%	94%	93%	94%	94%	94%	94%	94%
Obtained position	11	12	13	14	15	16	17	18	19	20
Percentage of urls with some missing	94%	94%	94%	95%	95%	94%	95%	95%	95%	95%
Obtained position	21	22	23	24	25	26	27	28	29	30
Percentage of urls with some missing	95%	95%	95%	96%	96%	95%	95%	94%	94%	95%

As we expect, not all the positions present the same number of missing values. Of course there are factors that are more complex to be obtained that others (does not represent the same difficulty to count how many times the keyword appears in the text, such as counting how many links point to the url that we are analyzing).
This can be clearly represented using a bar diagram that shows, first the frequency of observations we have for each one of the different positions, and next the frequency of observations but without missing values (see Figure 2).



Hist(AllNumeric$position, scale="frequency", breaks="Sturges", col="darkgray")



g<-na.omit(AllNumeric)
Hist(g$position, scale="frequency", breaks="Sturges", col="darkgray")

Figure 2. Frequency of observations for each position taking care of the missing values.

We cannot assume that the missing values are equally distributed over the different observations we have (MCAR) since clearly the first's positions have less missing values than others. This implies that it is needed to apply some kind of imputation.

In order to work with the data first it is needed to try to reduce the variables we have. To do so we calculate the correlation matrix according to Spearman and Pearson methods. The subjacent idea is to eliminate those variables that presents a high correlation. After performing those transformation we obtain a new matrix that presents only 111 variables.

2.2. Imputation

The imputation procedure is based on the substitution of the missing values by the nearest neighborhood. The method is the imputation based on random blocks, where we divide the data intro 20 blocks of randomly chosen rows and then impute missing using 9 nearest neighbors in its block.

```
# http://cran.r-project.org/web/packages/imputation/imputation.pdf
# kNNImpute {imputation}
source("UDFunctions.R")
r_factors2 <- knnimpute.by_random_blocks (ssr_factors,5)

knnimpute.by_random_blocks <- function(datos,nbloques) {

  ptmtot <- proc.time()
  n.filas <- nrow(datos)
  ns.filas <- ceiling(n.filas/nbloques)
  filas <- sample (1:n.filas,n.filas,replace=F)
  fi <- 1

  while (fi < n.filas){
    ptm <- proc.time()
    ff <- min(n.filas,fi+ns.filas-1)
    print (paste("Inputing rows", fi, "to", ff))
    sfilas <- filas[fi:ff]
    bloque <- datos[sfilas,]
    bloque <- bloque[,sapply(bloque, is.numeric)]
    res <- kNNImpute(bloque, k=9, verbose = FALSE)
    if ( fi == 1 ) {
      datos2 <- res$x
    }
    else {
      datos2 <- rbind(datos2, res$x)
    }
    fi <- ff+1
    print(proc.time() - ptm)
  }
  print(proc.time() - ptmtot)
  return(datos2[order(filas),])
}
```

3. Principal component analysis

Principal Component Analysis (PCA) is a statistical technique that is applicable to issues described with quantitative variables, as it is our case. The principal components analysis is formulated under the following assumptions:

1. Appropriate size of sample: it is recommended that the number of individuals or observed elements is greater than the number k of original variables.
2. They must avoid variables with the following characteristics:
    - Constant parameters.
    - Parameters with very little variance, i.e. virtually constant parameters.
    - Parameters that are linearly dependent on other parameters z = aX + bY
    - Parameters that do not depend on the other.
    - Parameters that change much.

After performing the initial reduction techniques and the imputation for the missing values we have a table containing 111 independent factors. Although the reduction in the number of factors, the main problem encountered in this dataset, statistically speaking, is the large number of factors we can use to try to explain the variable “position”.

In the following table you can see the percentage of variance explained by each component, and finally the cumulative explained variance.

TABLE

In this case, at first glance, it seems reasonable to stay with the first main 58 components, since with them is explained 90,37% of the variance, and bearing in mind that adding one more we win only 0.4%. Require such a large number of major components is important. This technique wants to reduce complex problems into 2-3 main components, in our case, since we need 58 principal components, makes us think that every one of the factors are important to Google's algorithm.

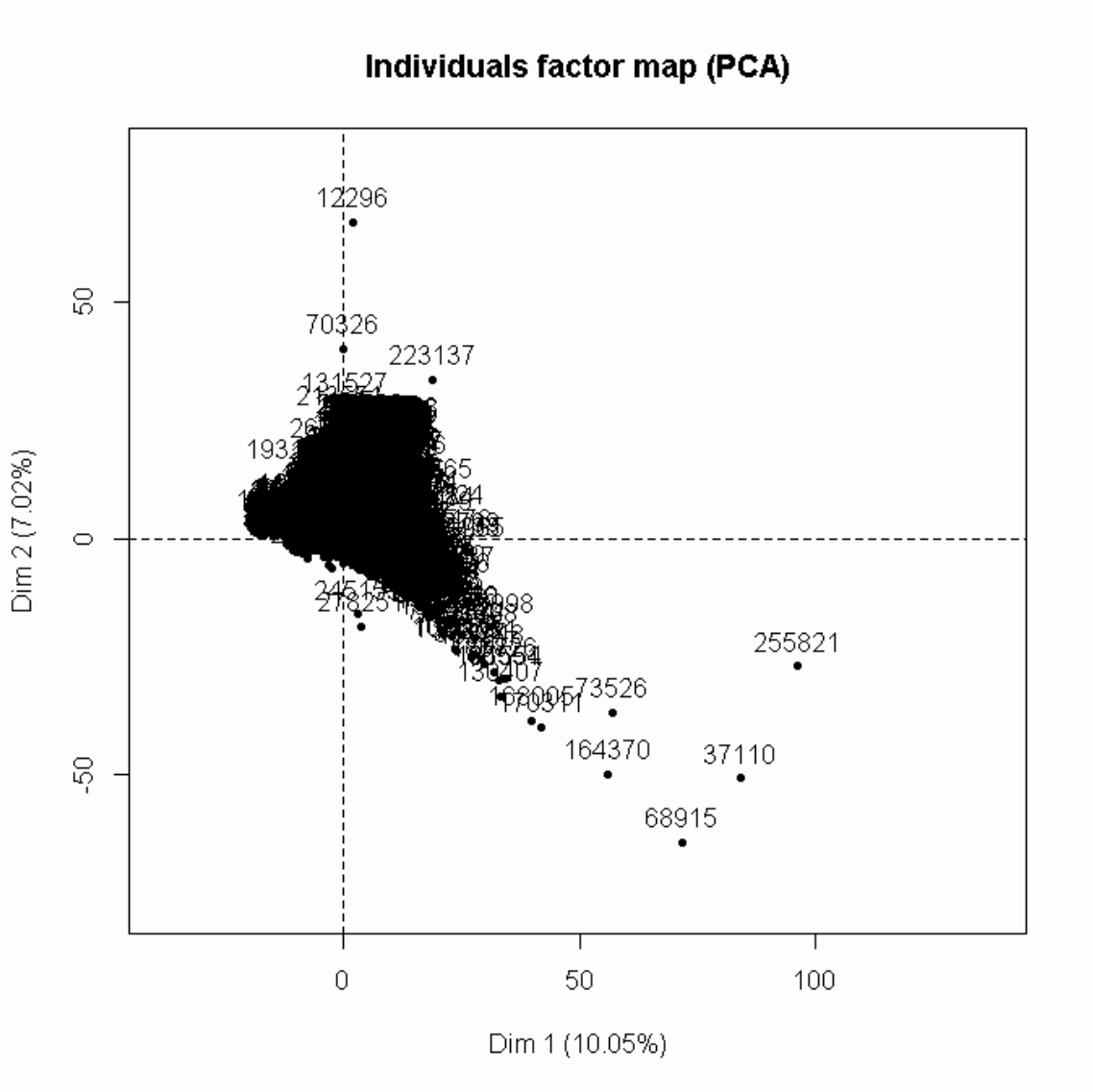


Figure 3. Individuals graph for the two main components.

We can see in Figure 3 the distribution of the observations based on the two first components. There is a no clear difference between the observations. We analyze in detail if there is any cluster in section 4.

It is interesting to note that there are several scattered individual observations. We refer to samples as the 255821, 37110, 68915, 164370, 73526, etc. that can be observed in Figure 3. The methodology of the principal components analysis suggests dismiss these observations and focus mostly on which are concentrated.

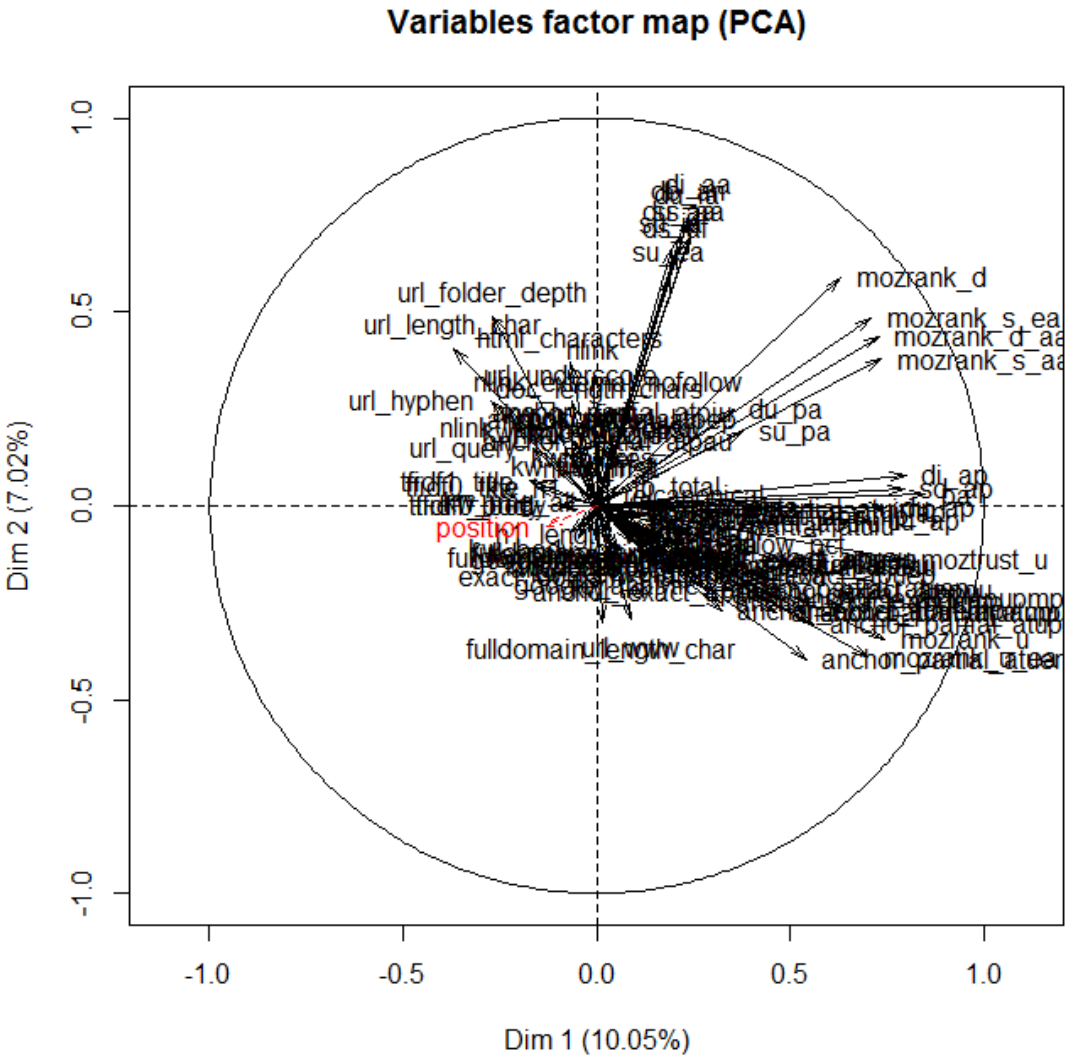


Figure 4. Principal Component Analysis ordination diagram.

Figure 4 shows component 1 is represented by the axis of abscissa (x axis) and the component 2 for the vertical axis (y axis). The analysis retained two axes that explained about 17% of the variation in the composition of the landscape. This low explanation of the variability makes that the conclusions must be taken very carefully. Looking in detail the graph we can see that the variable position (in red) have a totally horizontal left behavior. This means that its behavior is determined by the principal component 1 in negative direction (to the left). I.e., the lower the principal component 1 greater will be the position. Since we are interested position closer to 1, we are interested in the study of the behavior of positioning factors that minimize the main component 1, since these will reduce the final value of the position. Analyzing the eigenvectors for the two main axes of Figure 4 we can detect what are the main important factors for the position variable regarding these two

components. However the amount of variation explained is about 17%, meaning that there is a huge amount of the variation that is not explained by these two components.

The principal conclusion of the PCA analysis is that it seems that Google's algorithm is using all the retained factors we use on the PCA analysis. Going further we try to understand is some clusters exists that lead us to a detailed analysis of the algorithm's behavior.

4. Cluster analysis

To perform the cluster analysis we use Hierarchical Clustering on Principle Components (HCPC). Performs an agglomerative hierarchical clustering on results from a factor analysis. To determine the amount of clusters to retain we analyze Figure 5, a plot of the SSE against a series of sequential cluster levels. Figure 5 provides a useful graphical way to choose an appropriate cluster level, an appropriate cluster solution could be defined as the solution at which the reduction in SSE slows dramatically. This produces an "elbow" in the plot of SSE against cluster solutions. On Figure 5, there is an "elbow" at the 4 cluster solution suggesting that solutions >4 do not have a substantial impact on the total SSE.

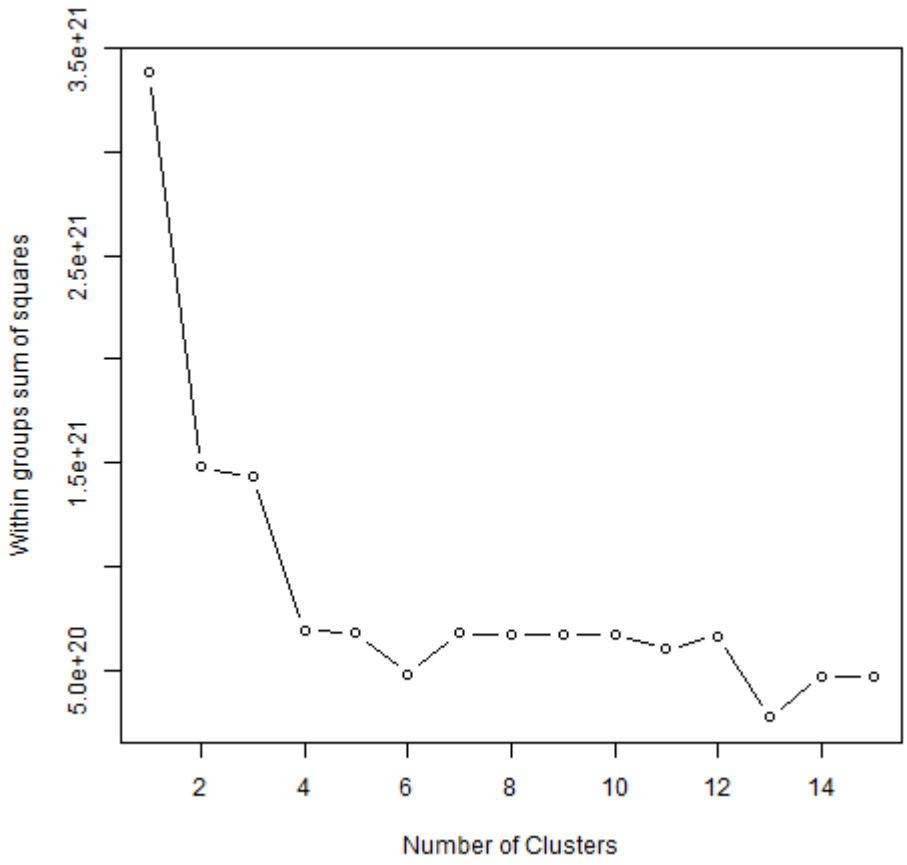


Figure 5. Number of clusters vs within groups sum of squares.

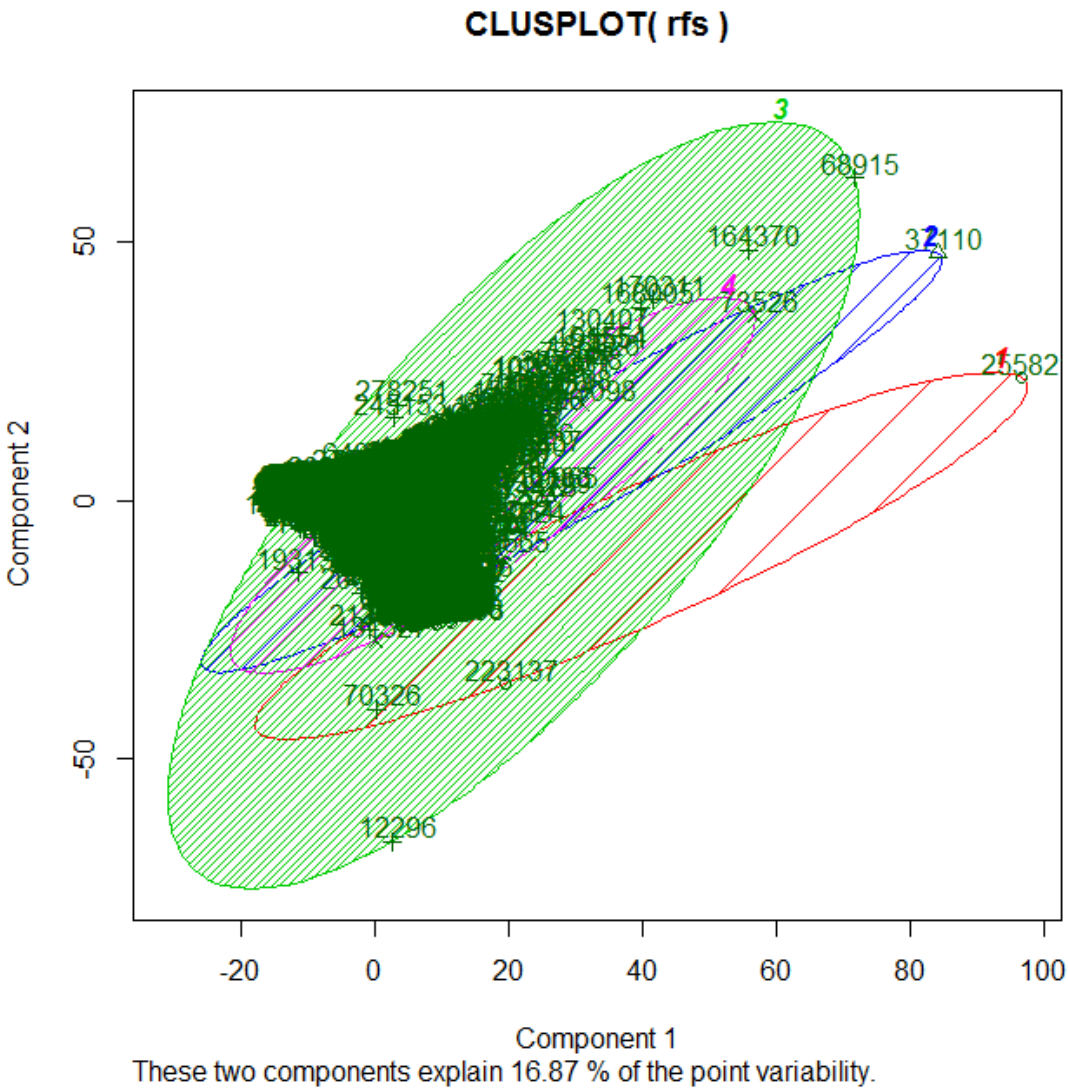


Figure 6. Cluster analysis.

We analyze the different proposed clusters, trying to understand first if exist a statistical difference between them. We use a non-parametric test since we cannot assure homoscedasticity or normality, in that case Kruskal-Wallist test. As we can see next, the test suggest that exist a statistical difference between the different groups, see also Figure 7.

```
kruskal.test(position ~ KMeans, data=rfs)

        Kruskal-Wallis rank sum test

data:  position by KMeans
Kruskal-Wallis chi-squared = 2628.566, df = 3, p-value < 2.2e-16
```

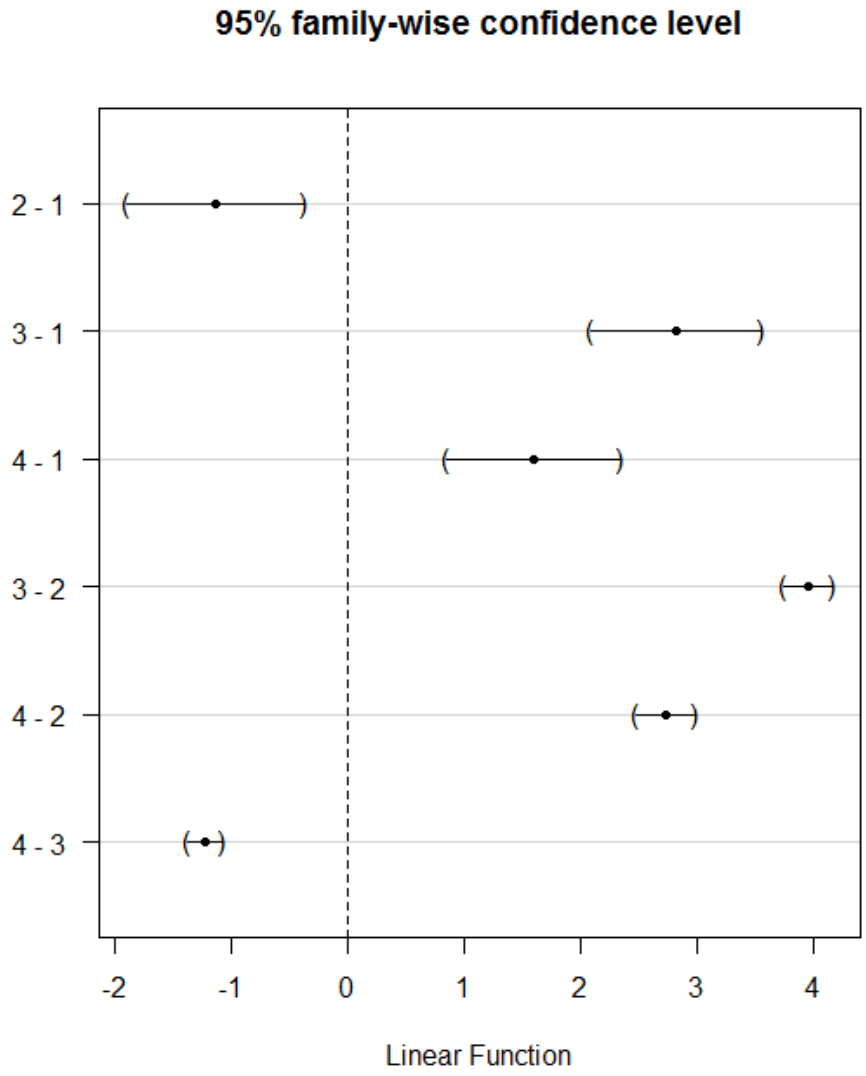


Figure 7. Exist a statistical difference between the clusters.

However, although this seems promising must be taken with caution. As is shown in the next table the amount of observations in each cluster is not homogeneous in any case.

Cluster	Observations
1	736
2	9386
3	251198
4	18087

Figure 8. Observation for each one of the different clusters proposed.

Performing a PCA for each one of the different groups do not improve the variability explained as we can see on next figure.

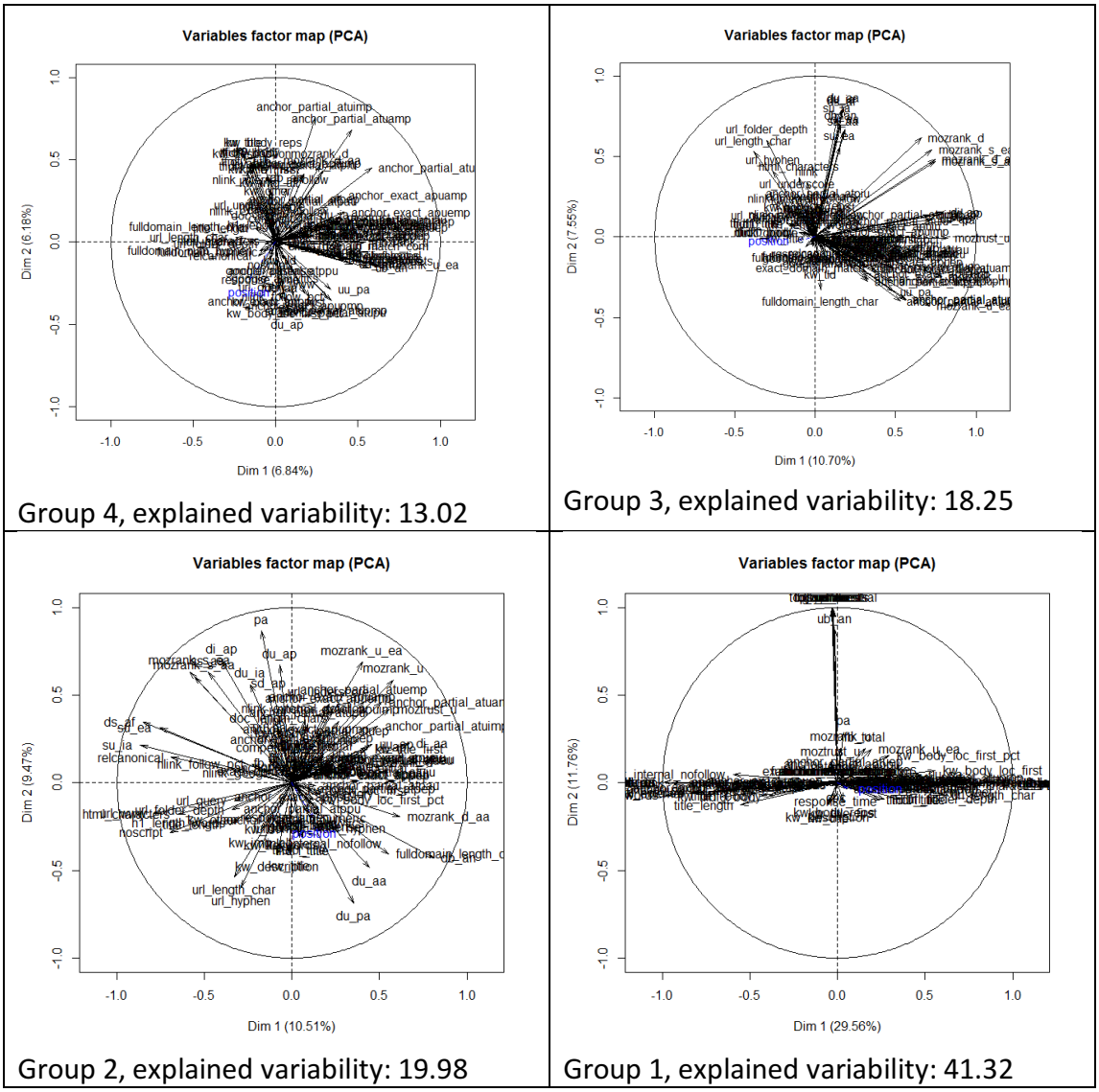


Figure 9. PCA for the different groups.

5. Concluding remarks

We present an analysis of Seomoz data in order to try to understand the key factors that determine a good position according Google's algorithm.

Although we perform a detailed analysis of the data, and a PCA to try to underline the existing relations between the different variables, the explained variability of the analysis is very low in all the cases. This implies that the analyzed data we use do not describe accurately the behavior of Google's algorithm leading to two possible interpretations.

1. The data that we have do not explain nothing regarding the position value.
2. The algorithm is using more or less equally all the different factors presented on the dataset.

Only in the group 1 it seems that the explained variability is higher than on the other groups.

More analysis are needed in order to understand the relation between the different factors proposed and the answer variable position.

6. Bibliography

Atrey, P. K., Ibrahim, H., Hossain, M. A., Ramanna, S., & El Saddik, A. (2012). Determining trust in media-rich websites using semantic similarity. *Multimedia Tools and Applications*, *60*(1), 69–96. doi:10.1007/s11042-011-0798-x

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation Bias Is Significant in Determining User Preference for Search Results-A User Study. *Journal of the American Society for Information Science and Technology*, *60*(1), 135–149. doi:10.1002/asi.20941

Beldona, S., Lin, K., & Chen, M. (2012). Hotel Trademarks in Organic Search: A Longitudinal Cross-National Study. *Journal of Travel Research*, *51*(2), 227–238. doi:10.1177/0047287511400612

Bifet, A., Castillo, C., Chirita, P., & Weber, I. (2005). An Analysis of Factors Used in a Search Engine's Ranking. Presented at the First International Workshop on Adversarial Information Retrieval on the Web. Retrieved from http://www.chato.cl/papers/bccw05_analysis_factors_search_engine_ranking.pdf

Bizer, C., & Oldakowski, R. (2004). Using context- and content-based trust policies on the semantic web. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 228–229). New York, NY, USA: ACM. doi:10.1145/1013367.1013409

Cutts, M. (2011). *Can you explain what Google means by "trust"?* Retrieved from http://www.youtube.com/watch?v=ALzSUeekQ2Q&feature=youtube_gdata_player

Cutts, M. (2012, April 24). Another step to reward high-quality sites. Retrieved from http://googlewebmastercentral.blogspot.com.es/2012/04/another-step-to-reward-high-quality.html

Cutts, M. (2013, February 22). A reminder about selling links that pass PageRank. Retrieved April 24, 2013, from http://googlewebmastercentral.blogspot.com.es/2013/02/a-reminder-about-selling-links.html

Czarnowski, I., & Jędrzejowicz, P. (2010). An Agent-Based Simulated Annealing Algorithm for Data Reduction. In P. Jędrzejowicz, N. T. Nguyen, R. J. Howlet, & L. C. Jain (Eds.), *Agent and Multi-Agent Systems: Technologies and Applications* (pp. 130–139). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-13541-5_14

Fortunato, S., Boguna, M., Flammini, A., & Menczer, F. (2005). How to make the top ten: Approximating PageRank from in-degree. *arXiv:cs/0511016*. Retrieved from http://arxiv.org/abs/cs/0511016

Geyskens, I., Gielens, K., & Dekimpe, M. G. (2002). The market valuation of Internet channel additions. *Journal of Marketing*, *66*, 102–119. doi:10.1509/jmkg.66.2.102.18478

Goodwin, D. (2013, April 3). UK Flower Site Banned from Google for Advertorial Links Sees Rankings Restored. Retrieved April 24, 2013, from http://searchenginewatch.com/article/2252233/UK-Flower-Site-Banned-from-Google-for-Advertorial-Links-Sees-Rankings-Restored

Google. (2013a). Google Basics - Webmaster Tools Help. Retrieved April 16, 2013, from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=70897&ctx=sibling#1

Google. (2013b). Webmaster Guidelines - Webmaster Tools Help. Retrieved April 24, 2013, from http://support.google.com/webmasters/bin/answer.py?hl=en&answer=35769

Gyongyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating Web Spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*. Toronto, Canada. Retrieved from http://ilpubs.stanford.edu:8090/770/

Hariri, N. (2011). Relevance ranking on Google Are top ranked results really considered more relevant by the users? *Online Information Review*, *35*(4), 598–610. doi:10.1108/14624521111161954

Haveliwala, T. H. (2002). Topic-Sensitive PageRank. In *To appear in WWW-2002*. Honolulu, Hawaii. Retrieved from http://ilpubs.stanford.edu:8090/573/

Hoechstoetter, N., & Lewandowski, D. (2009). What users see - Structures in search engine results pages. *Information Sciences*, *179*(12), 1796–1812. doi:10.1016/j.ins.2009.01.028

Jansen, B. J., Brown, A., & Resnick, M. (2007). Factors relating to the decision to click on a sponsored link. *Decision Support Systems*, *44*(1), 46–59. doi:10.1016/j.dss.2007.02.009

Jansen, B. J., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, *52*(3), 235–246. doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.3.CO;2-6

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263. doi:10.1016/j.ipm.2004.10.007

Jansen, B. J., Zhang, M., & Schultz, C. D. (2009). Brand and its Effect on User Perception of Search Engine Performance. *Journal of the American Society for Information Science and Technology*, *60*(8), 1572–1595. doi:10.1002/asi.21081

Khaki-Sedigh, A., & Roudaki, M. (2003). Identification of the dynamics of the google's ranking algorithm. In *In 13th IFAC Symposium On System Identification*.

Killoran, J. B. (2009). Targeting an Audience of Robots: Search Engines and the Marketing of Technical Communication Business Websites. *IEEE Transactions on Professional Communication*, *52*(3), 254–271. doi:10.1109/TPC.2009.2025309

Killoran, J. B. (2010). Writing for Robots: Search Engine Optimization of Technical

Communication Business Web Sites. *Technical Communication*, *57*(2), 161–181.

Killoran, J. B. (2013). How to Use Search Engine Optimization Techniques to Increase Website Visibility. *Ieee Transactions on Professional Communication*, *56*(1), 50–66. doi:10.1109/TPC.2012.2237255

Kowalski, P., & Król, D. (2010). An Approach to Evaluate the Impact of Web Traffic in Web Positioning. In P. Jędrzejowicz, N. T. Nguyen, R. J. Howlet, & L. C. Jain (Eds.), *Agent and Multi-Agent Systems: Technologies and Applications* (pp. 380–389). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-13541-5_39

Lam, S. (2001). The Overview of Web Search Engines. *University of Waterloo*. Retrieved from http://db.uwaterloo.ca/~tozsu/courses/cs748t/surveys/sunny.pdf

Lee, S. J., Kang, H. W., & Chung, J. (2012). An Empirical Analysis on the Marketing Performance of Korean Exporting Companies by Major Industrial Classification using Search Engine Optimization. *Journal of Korea Trade*, *16*, 79–110. Retrieved from ://WOS:000301021700005

Liang, S. F., Devlin, S., & Tait, J. (2006). Evaluating web search result summaries. In M. Lalmas, A. MacFarlane, S. Ruger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in Information Retrieval* (Vol. 3936, pp. 96–106). Berlin: Springer-Verlag Berlin.

Lorigo, L., Pan, B., Hernbrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence fo task and gender on search and evaluation behavior using Google. *Information Processing & Management*, *42*(4), 1123–1131. doi:10.1016/j.ipm.2005.10.001

Malaga, R. A. (2008). Worst Practices in Search Engine Optimization. *Communications of the Acm*, *51*(12), 147–150. doi:10.1145/1409360.1409388

Moran, M. (2013a). *Search engine marketing, inc.: driving search traffic to your company's web site.* [S.l.]: Ibm Press.

Moran, M. (2013b). *Search engine marketing, inc.: driving search traffic to your company's web site.* [S.l.]: Ibm Press.

Moran, M., & Hunt, B. (2006). Search Engine Marketing, Inc. driving search traffic to your company's web site. Retrieved March 19, 2013, from http://www.books24x7.com/marc.asp?bookid=12239

Pan, B., Xiang, Z., Law, R., & Fesenmaier, D. R. (2010). The Dynamics of Search Engine Marketing for Tourist Destinations. *Journal of Travel Research*, *50*(4), 365–377. doi:10.1177/0047287510369558

Papatla, P., & Liu, F. (2009). Google or BizRate? How search engines and comparison sites affect unplanned choices of online retailers. *Journal of Business Research*, *62*, 1039–1045. doi:10.1016/j.jbusres.2008.10.018

Paraskevas, A., Katsogridakis, I., Law, R., & Buhalis, D. (2011). Search Engine Marketing: Transforming Search Engines into Hotel Distribution Channels. *Cornell Hospitality Quarterly*, *52*, 200–208. doi:10.1177/1938965510395016

Rimbach, F., Dannenberg, M., & Bleimann, U. (2007). Page ranking and topic-sensitive page

ranking: micro-changes and macro-impact. *Internet Research*, *17*(1), 38–48. doi:10.1108/10662240710730489

Sabate, F., Canabate, A., Velarde-Iturralde, M.-A., & Grinon-Barcelo, R. (2010). Use of internet promotion strategies by the Spanish travel agencies. *Profesional De La Informacion*, *19*, 149–159. doi:10.3145/epi.2010.mar.05

Sen, R. (2005). Optimal search engine marketing strategy. *International Journal of Electronic Commerce*, *10*, 9–25. Retrieved from ://WOS:000233211600002

SEOmoz. (2013a). Google Algorithm Change History. Retrieved April 23, 2013, from http://www.seomoz.org/google-algorithm-change

SEOmoz. (2013b). MozTrust - SEO Best Practices. Retrieved April 25, 2013, from http://www.seomoz.org/learn-seo/moztrust

SEOmoz. (2011). 2011 Search Engine Ranking Factors. Retrieved April 11, 2013, from http://www.seomoz.org/article/search-ranking-factors#metrics

Silverman, D. (2012). IAB Internet Advertising Revenue Report. 2011 Full Year Results. Retrieved from http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2011.pdf

Singhal, A. (2011, May 6). More guidance on building high-quality sites [Blog]. Retrieved from http://googlewebmastercentral.blogspot.com.es/2011/05/more-guidance-on-building-high-quality.html

StatCounter. (2013, January 5). Top 5 Search Engines from Jan to Apr 2013 | StatCounter Global Stats. Retrieved April 30, 2013, from http://gs.statcounter.com/#search_engine-ww-monthly-201301-201304-bar

Suchanek, P. (2010). THE FUNDAMENTALS OF A PROSPEROUS E-SHOP IN CONNECTION TO SEARCH ENGINE OPTIMIZATION. *E & M Ekonomie a Management*, *13*, 92–103. Retrieved from ://WOS:000283638300009

Sullivan, D. (2002). The Mixed Message Of Paid Inclusion. Retrieved January 23, 2013, from http://searchenginewatch.com/article/2048435/The-Mixed-Message-Of-Paid-Inclusion

Wall, A. (2005, February 7). TrustRank Algorithm. Retrieved April 25, 2013, from http://www.seobook.com/archives/000661.shtml

Weideman, M. (2007). Use of Ethical SEO Methodologies to Achieve Top Rankings in Top Search Engines. In *Proceedings of the 2007 Computer Science and IT Education Conference* (pp. 717–727). Republic of Mauritius (Indian Ocean).