A Search Engine Optimization Recommender System

Christian D. Hoyos, Juan C. Duque, Andrés F. Barco, Élise Vareilles

A Search Engine Optimization Recommender System

Christian D. Hoyos [1] and **Juan C. Duque**[1] and **Andrés F. Barco**[2] and **Élise Vareilles**[3]

Abstract. Search Engine Optimization reefers to the process of improving the position of a given website in a web search engine results. This is typically done by adding a set of parameters and metadata to the hypertext files of the website. As nowadays the majority of the web-content creators are non-experts, automation of the search engine optimization process becomes a necessity. On this regard, this paper presents a recommender system to improve search engine optimization based on the site's content and creator's preferences. It exploits text analysis for labels and tags, artificial intelligence for deducing content intention and topics, and case-based reasoning for generating recommendations of parameters and metadata. Recommendations are given in natural language using a predefined set of sentences.

1 Introduction

Normally, web content creators require their websites to be easily found by content consumers through search engines [6]. They do so by setting parameters and adding metadata to the hypertext source files of the websites. These parameters and metadata allow the algorithms of the search engines to index and retrieve data of millions of websites in an efficient way [7]. For instance, parameters about the intention of the website allow to classify content and metadata stating the location is useful to customize content or restrict access. Further, this information makes possible for the search engine to rank the results of a query by priority. As reported by Chitika [5], configuring websites for correct indexing is a key element of their success. This configuration of values is called Search Engine Optimization (SEO).

Now, although every website is implemented following a standard, namely HTML, there is no standard for web page ranking as each search engine (Google, Yahoo, Bing, etc) implements its own ranking system. This implies that improving the indexing position of a website requires an expert on both the content as well as on the search engine ranking system.

On this regard, this paper proposes an expert recommendation system in charge of performing SEO for a given web page[4] targeting the Google search engine. It uses artificial intelligence to deduce the intention and content topic of the web page, it uses text analysis over labels and tags in order to classify and comparison, and it uses case-based reasoning to provide recommendations for improving SEO on the web page.

The documents is structured as follows. The overall behavior of the system, and its architecture, are presented in Section 2. Each of the modules of the system are described in Section 3. An experimental test and its results are shown in Section 4. Conclusions are presented in Section 5.

[1] Universidad de San Buenaventura Cali. Santiago de Cali, Colombia. email: {christian, juan.duque}@gmail.com
[2] Universidad Santiago de Cali. Santiago de Cali, Colombia. email: anfelbar@usc.edu.co
[3] Université de Toulouse, Mines Albi. Albi, France. email: elise.vareilles@mines-albi.fr
[4] This means it analyzes each web page individually.

2 Overview

To provide recommendations for indexation of a web page, aspects such as content topic, keywords, intention of the (authors') web page, metadata, related web pages and the specific raking system of the search engine, should be taken into account. These aspects allow the expert system to understand the website communication goals and to create recommendations that respect the search engine implementation. The expert system proposed here tries to unveil the previous aspects using three modules in charge of analysis and one module in charge of recommendation generation (see Figure 1).

The systems receives three inputs, two of which are optional. The first input of the system is either an HTML source file or an hyperlink (URL to an HTML). If the HTML contains scripts or CSS definitions, they are ignore are they not provide useful information for the indexation. Hyperlinks should be accessible from the web. The second input is the topic of the web page, which is an optional value. The last input value is the intention of the web page and it is as well optional. It is worth noticing that having explicitly defined topic and intention will help the system's accuracy and performance (no topic and intention processing). Having the inputs, the system executes the following steps and throws as output a web page score and its recommendations.

First, the web page is analyzed using text analysis over the HTML source code. The analysis throws an score depending on the presence or absence of 22 of the more important factors for indexation according to Google [2, 5]. These factors add positive values to the score when present and negative values when not. This is the first source of knowledge to build a recommendation of a web page.

Once the text analysis is done, a topic and intention analysis is performed using the IBM Watson system (a state-of-the-art artificial intelligent API) [9]. The topic and intention are useful in two ways. At the one hand, they allows to classify the content of the web page. And, on the other hand, they are basis a case-based reasoning recommendation executed in the last step.

Next, using the obtained topic and intention as keywords, the system performs a search query in the Google search engine and retrieves the first 10 pages from the result. It then proceeds by analyzing each web page in the aim of extracting key values, such as keywords and metadata, that made those pages the 10 first ranked pages of Google. This is an implementation of case-based reasoning [8] and are the second source of knowledge to build a recommendation of a web page.

Finally, the system builds a recommendation using `HTML` code and natural language [4] using predefined sentences. They are based on

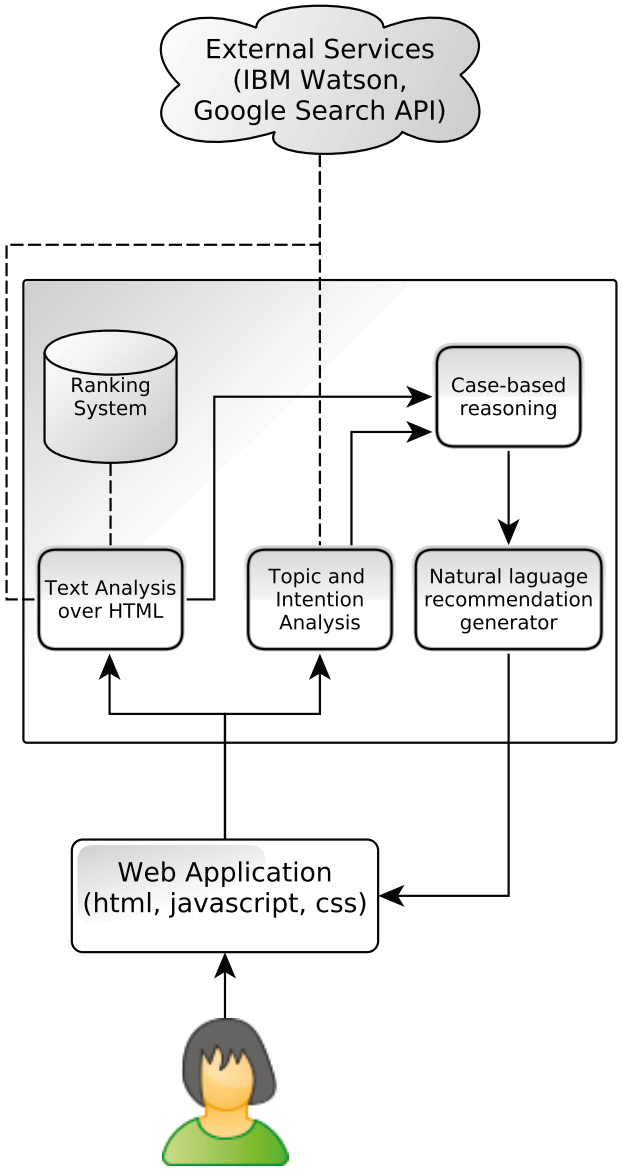

Figure 1. Recommendation System Architecture.

the identified negative evaluated factors (e.g., missing tags) and the extracted data from the first 10 pages (e.g., new keywords).

3 System's Core

The recommendation system is divided in four modules.

3.1 Module 1: HTML analysis

This module focuses in labels and metadata of the web page HTML source files. In particular, it looks for specific information that is related with the Google ranking system and 22 key aspects in specific labels as `<meta name=...>`. These aspects include keywords definition, char-set codification, description of web page, copyright, content duplication and broken links, among others. Each factor has associated a positive value if included in the source file and a negative value if not. The Table 1 present some of the key aspects and its respective values.

Label	Description	Benefit	Penalty
F1	User of keywords in tag `title`.	13,5	-16,8
F2	Connection among keywords (interrelated)	13,5	-16,8
F3	Low density on keywords (not too many)	10,5	-16,8
F4	Description in tag `meta` with a maximum of 200 words.	10,5	-16,8
F5	Excessive use of `meta` and `alt` tags.	10,5	-16,8
F6	Definition codification in tag `char-set`.	13,5	-12,6
F7	Avoid the use of tag `refresh`	7,5	-16,8
F8	Use of tag `alt` in `<img>` and `<input>`	12	-12,6
F9	No broken URLs in source file	13,5	-10,5
F10	Use of tag `H` (h1, h2, h3)	10,5	-12,6
F11	Exceeding maximum number of characters in tag `title`	6	-14,7
F12	Use of tag `keyword` with maximum of 200 characters.	12	-8,4
F13	Percentage (between 5 and 20) of keywords in text	10,5	-8,4
F14	Hyperlinks to pages of the same website	13,5	-4,2
F15	Content strongly connected to the web page topic and keywords	10,6	-6,3
F16	Duplicated content.	10,5	-6,3
F17	Use of `strong`, `bold` and `italic` for fonts.	12	-4,2
F18	Use of `cache-control` tag.	9	0
F19	Keywords in URL.	6	-6,3
F20	Use of keywords in numbered lists.	7,5	-4,2
F21	Use of tag `author`.	3	-2,1
F22	Definition of tag `copyright`	3	-2,1

Table 1. Evaluated factors and scoring.

Note: It is important to know that one of the most important factor in the Google search engine is the value determined by the PageRank algorithm [1]. This algorithm takes into account the number and quality of other web pages pointing to the web page in reference. Simply put, the more pages on the web point at the referenced page the better. More points are given if the other web page is high ranked. This works as a kind of endorsement. The PageRank is not included in the recommendation system analysis as it is not based in HTML tags and metadata.

3.2 Module 2: Intention and topic deduction

The intention and topic is deduced from the content, meaning that only the text within the labels `<body> ... </body>` are analyzed. Both intention and topic are deduced using the IBM Watson system through its public API only if no user input is given. Watson is, en essence, an on-line system that exploits several techniques from artificial intelligence to provide services as speech to text, natural language understanding and query answer system, emotion and sentiment analysis, translator and visual recognition [3, 9].

The topic and intention are deduced by Watson using Natural Language Understanding/Classification for the analysis of text. In the case of the topic, classification is done through a set of categories, concepts and keywords. In case of the intention the system classifies according to how positive or negative is the web page. Then the analysis assigns one of the following labels to the page: Very Positive, Positive, Neutral, Negative and Very Negative. Each of these labels are connected to numerical values thrown by Watson, as presented in Table 2.

Label	Min	Max
Very Positive	0.6	1
Positive	02	0.6
Neutral	-0.2	0.2
Negative	-0.6	-0.2
Very Negative	-1	-0.6

Table 2. Table with this

3.3 Module 3: Case-based reasoning

The set of categories, concepts, keywords and the intention are used for constructing a search query in the aim of obtaining similar web pages. The main idea is to extract the parameters used by top ranked web pages, the first 10 pages in Google's search engine, that address the same topic and has the same intention. Potentially, those 10 pages include data in their HTML files that made them the first ranked by the search engine. Arguably, using the same or similar parameters (as new keywords or tags), will help to improve indexation of other pages. For instance, adding keywords that were not previously included in the web page but that are a common for most of the 10 retrieved pages.

Note: The system only retrieves the first 10 pages for two reasons. At the one hand, according to the literature, the probability of user access to a web page ranked after 10th position is around 1% [5]. Thus, the system obtains only those web pages that are likely to have high user access rate. And, at the other hand, the analysis of more pages may reduce its efficiency. Consider that each of the 10 pages is analyzed using the same techniques. Ergo, the process, plus comparison, must be executed 11 times, which is time consuming.

3.4 Module 4: Natural language recommendation

Recommendation are build with structured predefined sentences of the form: target factor + recommendation over factor + explanation of recommendation + example in HTML. Each recommendation is classified into four categories in according with its importance:

- Black: Critical recommendation to be applied for basic indexation in Google search engine.
- Red: Not following the recommendation may significantly affect the position of the web page in the results.
- Yellow: Not following the recommendation may moderately affect the position of the web page in the results.
- Blue: Not following the recommendation may minimally affect the position of the web page in the results.

4 Test

Two type of tests have been made; tests using public web pages and tests using an authors' web page.

4.1 Public websites tests

In these tests, five topics have been chosen and the following five queries have been designed.

1. Football soccer critic.
2. Mediterranean food.
3. Vaccines for cats.
4. Contamination of the Oceans.
5. Renewable energies.

The first three results of each query have been feed to the system with automatic execution. Table 3 shows the number of recommendations of each found page.

Query	Index	# Reco	Score
Football critic	1	43	191
	2	18	215
	3	63	292
Mediterranean food	1	32	494
	2	14	111
	3	33	114
Vaccinations for cats	1	25	306
	2	16	439
	3	23	171
Contamination of the Oceans	1	36	222
	2	17	67
	3	48	50
Renewable energies	1	66	600
	2	65	223
	3	10	90

Table 3. Test results with five queries.

From the results we conclude the following.

- There is no direct relation between the number of recommendations and the score of the web page. Indeed, it depends on what factor is being recommended and its impact on the score. For instance, a web page may have few recommendations on factors, but one of the factor is being repeated within the page ergo reducing the score significantly.
- Of the five queries, three of them show tendency of score decreasing, which is expected. The first pages of the other two queries do not have high score but may be affected by other factor not taken into account, mainly traffic and results of PageRank algorithm.
- The best ranked pages are part of sites like Wikipedia. In fact, two of the found web pages are from Wikipedia and are the top ranked pages. This is mostly due to the many external and self-reference hyperlinks of the site.

4.2 Authors' designed web page

For these tests, a web page created by the authors is feed to the system in three different round. Recommendations (from rounds one and two) are implemented before the next round (rounds two and three). The designed page is a basic HTML file, without styles or scripts,

used to show the improvement of a given web page through the system's recommendations. The title of the web page is the "The fall of JQuery", and addresses the descend of developers using `JQuery`. Figure 2 show the recommendations of the first round (in Spanish[5]) with different colors for their importance. As an example of the result, first line of recommendation states *"You should use labels h1, h2, h3...h6 more often, as they help defining the importance of conent within the page"*. Table 4 presents the results of the three rounds of execution.

Round	# Recommendations	Score
1	13	-116
2	5	100
3	0	167

Table 4. Follow-up of authors' web page.

The values thrown by the system in the three rounds show an evolution of the web page through the recommendations. As expected, for a web page with no external links referencing at it, the system assigns low score and several recommendations for the first run, augmenting score and decreasing recommendations. Bear in mind that the number of recommendations is lower than the tests in the previous sections given that the content of the designed web page is not as big and does not have as many links as the other pages.

5 Conclusions

Although the internals of web search engines are very similar, each of them implements different ranking system for indexing web pages. In consequence, the identification of factors that are included in the ranking systems, and its tuning by means of hypertext (metadata), is critical for the success of a given web page. In this context, tags, topic and intention are relevant for recommending changes in the aim of improving results position.

This paper proposed a recommender system for improving the search optimization of a web page in the Google's search engine. The system evaluates 22 main factors used by Google search engine to classify the web pages (ranking them). The system represents a positive contribution because:

- Basic and fundamental factors are handled so that the search engine can identify the content and structure of the web page.
- Each recommendation explains with details and examples, and in natural language, how the improvement of a factor in the website can be made.
- An user without much experience in SEO can make use of the recommendation system as it is intuitive.
- Recommendations are different for each factor and each web page (customized recommendations).
- The analysis and recommendations are made based on the top 10 bests indexed sites in Google, that deal with the same topic and intention (instance of case-based reasoning).

REFERENCES

[1] Monica Bianchini, Marco Gori, and Franco Scarselli, 'Inside pagerank', *ACM Trans. Internet Technol.*, **5**(1), 92–128, (February 2005).

[2] Pablo Fernández, 'Google's pagerank and beyond: The science of search engine rankings', *The Mathematical Intelligencer*, **30**(1), 68–69, (Mar 2008).

[3] D. A. Ferrucci, 'Introduction to "This is Watson"', *IBM Journal of Research and Development*, **56**(3.4), 1:1–1:15, (May 2012).

[4] Chowdhury G., 'Natural language processing', *Annual Review of Information Science and Technology*, **37**(1), 51–89, (2003).

[5] Chitika Insights. The value of google result positioning. `http://info.chitika.com/uploads/4/9/2/1/49215843/chitikainsights-valueofgoogleresultspositioning.pdf`, cited June 2019.

[6] J. B. Killoran, 'How to use search engine optimization techniques to increase website visibility', *IEEE Transactions on Professional Communication*, **56**(1), 50–66, (March 2013).

[7] Atanas Kiryakov, Borislav Popov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, and Miroslav Goranov, 'Semantic annotation, indexing, and retrieval', in *The Semantic Web - ISWC 2003*, eds., Dieter Fensel, Katia Sycara, and John Mylopoulos, pp. 484–499, Berlin, Heidelberg, (2003). Springer Berlin Heidelberg.

[8] J. Kolodner, *Case-Based Reasoning*, Elsevier Science, 2014.

[9] Punkaj Vohra, 'The new era of watson computing this article introduces cognitive computing using ibm watson and how to leverage cognitive computing with ecm centric solutions', *IBM Developer Works*, (02 2014).

[5] The system interface is in Spanish as it is being used in a multimedia engineering program in Colombia.

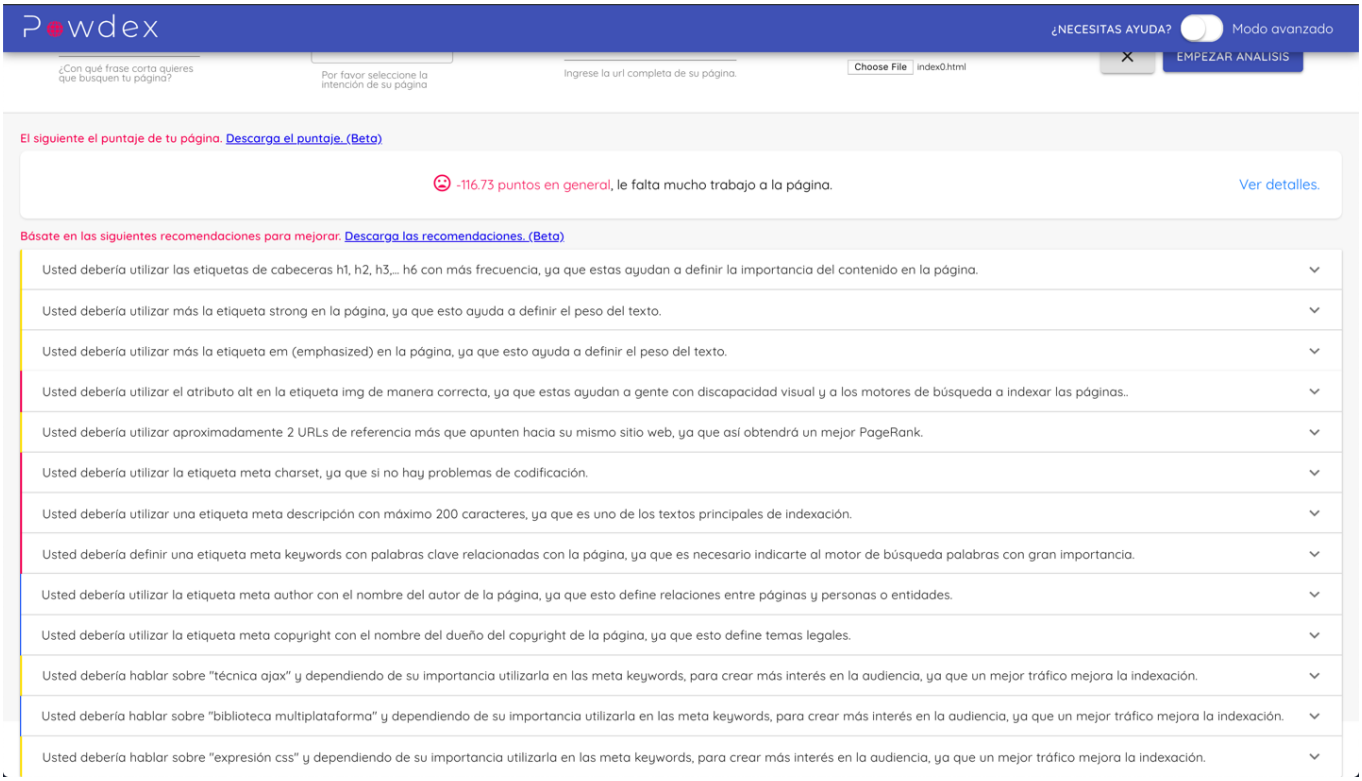

Figure 2. Recommendations for first round to "The fall of JQuery" web page..