Artículo original

User preferences, query and document categorization, three important variables in relevance calculation

Preferencias del usuario, categorización de consultas y documentos, tres variables importantes en el cálculo de relevancia

Paúl Rodríguez Leyva[1] https://orcid.org/0000-0002-2949-0766
Hubert Viltres Sala[1] https://orcid.org/0000-0002-5116-3665
Juan Pedro Febles1* https://orcid.org/0000-0003-3126-7667
Vivian Estrada Sentí[1] https://orcid.org/0000-0002-7513-7891
Yennifer Delgado Mesa[1] https://orcid.org/0000-0002-2949-0766

[1]Universidad de las Ciencias Informáticas. La Habana, Cuba.

*Autor para la correspondencia: febles@uci.cu

ABSTRACT

The quality of an information retrieval system depends largely on the satisfaction degree of users with the results obtained when executing a query, so it is essential to design processes that store the preferences patterns of each of them and vary the way in which the results are shown taking into account the specific characteristics of each user. The objective of this article was to present an algorithm for calculating the relevance of the documents provided to users, which used the variables: the user's search profile, the category of the documents and the category of the query as parameters, to customize the results provided by the search engine to the users. In addition, it used as impulse factors the degree of predominance of a search category in the user's profile and the categories to which the document belongs. To validate the algorithm, precision and recall metrics were applied to check that the results obtained are relevant to users.

Key words: Relevance; information retrieval; user profile; preferences; algorithm.

RESUMEN

La calidad de un sistema de recuperación de información depende en gran medida del grado de satisfacción de los usuarios en cuanto a los resultados obtenidos al realizar una consulta. Para obtener resultados de búsquedas relevantes es esencial diseñar procesos que almacenen los patrones de preferencias de cada usuario. Este estudio tuvo como objetivo presentar un algoritmo para el cálculo de la relevancia de los documentos brindados. El algoritmo utilizó como parámetros las siguientes variables: perfil de búsqueda del usuario, categoría de los documentos y categoría de la consulta para personalizar los resultados brindados mediante el motor de búsqueda. Además, utilizó como factores de impulso el grado de predominio de una categoría de búsqueda en el perfil del usuario y en las categorías a las que pertenece el documento. Para la validación del modelo se aplicaron las métricas de precisión y exhaustividad que permitieron comprobar que los resultados obtenidos son relevantes para los consumidores de la información.

Palabras clave: Relevancia; recuperación de información; perfil de usuario; preferencias; algoritmo.

Recibido: 10/04/2020

Aceptado: 14/04/2020

Introduction

In an interview with *Baeza Yates* by *Marcos*[1] he states that the main insufficiency that persists in search engines is trying to understand the intention after the search, that is, what is the informational need of people and customize their searches to that task. This involves predicting the intention and adapting the interface to the whole task. This is one of the challenges that IRSs (Information retrieval Systems) face at present, each user has characteristics that make them distinctive in the IR (Information Retrieval) process, so the personalization of the results they receive as answers to their search queries should be the essence of the internal functioning of a search engine.

These aspects are valued by companies such as Google that recognizes how the update of their algorithm for the calculation of relevance (Maccabees) takes into account the update focuses on the quality of the content, the links pointing to the site and the experience of the user.[2]

This change represents a revolution in the way in which the websites are positioned in this IRS, since the quality of the content given to the users and the relationship it has with their interests are prioritized. Although search engines can offer millions of results for a single search, in reality this is not important. From the point of view of the final user, it is indifferent that, for a key word, the search engine has either a few tens or several million results, since it will only examine the firsts.[3] For this reason, it is important to develop relevance calculation algorithms so that a response which satisfies the user's query is provided in the minimum number of results. In order to get a better response, in this work we propose an algorithm which includes and integrates user preferences and document categories.

This paper is divided into a related work section, where the state of the art about the calculation of relevance is exposed; the section methods, where it is explained an algorithm which integrates document categories; and user profile to calculate the relevance of a document, the results section, where the results obtained in an experiment with the algorithm are shown.

Related work

The personalization models in search engines have the goal of bringing to the users personalized results, decreasing the amount of irrelevant documents.[4] Several researches have shown how they are used as sources to define user preferences, queries, documents consulted, user browsing history, interaction with social networks, the nature of documents, or concepts associated with documents, among others; recognizing as impact factors the history of queries inserted by users and the nature of the documents in the collection. The most referred techniques in the previously discussed researches are focused on re-ranking methods, Bayesian classifiers, ontology design, terms frequency, agent technology, methods based on inferences, greedy algorithms, stemming algorithms, among others.[4,5,6,7,8,9,10,11,12]

Despite the fact that positive results have been demonstrated in some of these methods, it is difficult to find an algorithm explained in detail to integrate the user profile variables and the nature of the documents in the relevance calculation process.

According to *Fransson*,[13] in the case of Google, its administrators say they use more than 200 variables in their algorithm; however, they are not explained in depth in order to develop similar forms of calculation. Some of the most important updates since 2011 are the following table:[14]

Table 1 - Updates of the Google algorithm

Name	Panda	Penguin	Hummingbird	Pigeon	Mobile	RankBrain	Possum	Fred
Date	February 2011	April 2012	August 2013	July 2014	April 2015	October 2015	September 2016	March 2017

Source: 8 major Google algorithm updates, explained. Available from: https://searchengineland.com/8-major-google-algorithm-updates-explained-282627

In the limited bibliography referring to the specificities of the Google algorithm, it is possible to recognize the importance played by the calculation of the relevance of the user's profile definition and the nature of the query, but there is no specific information on how these variables are integrated in the implementation of their algorithm.[15,16,17]

Methods

The proposal presented allows to solve the problem described above, by designing and implementing an algorithm[1] that integrates the categories of the user's search profile (USP) and that of the documents (DC) to retrieve relevant and personalized information. These categories can be defined prior to the execution of the algorithm, usually calculated statically. A dynamic version could be defined; however, it is not the goal of the present work. The USP is defined as a set of pairs category-index, where indexes are selected by users in the interface registration and in the thematic search interface, in both interfaces the form to select the index are the same (Fig. 1).

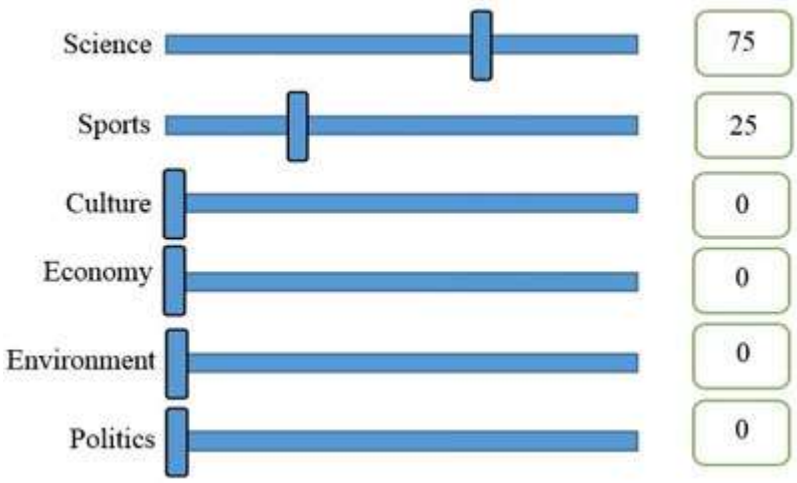

Fig. 1 - Form to define de index of relevance for each category.

The query categories (QC) is defined as the relation of the categories to which the query belongs with its percentage of predominance after executing the process of categorization of the query. In the example shown in table 2, the query is categorized with a 60% predominance belonging to the environment category, 20% to politics and 20% to culture. To store the QC, is proposed a matrix that relates the categories to which the query belongs and their membership percentages.

Table 2 - Query categories

Category	Environment	Politics	Culture
Index	0,6	0,2	0,2

The expression (1) is applied to two scenarios in which the relevance $R(u,q,d)$ differs in its values:

$$R(u,q,d) = \alpha\ SCD(q,d) + \beta\ RCD(q,d,ts) + \gamma\ RPUD(q,d,rp) \quad (1)$$

where *ts* is the relation category-index defined in the thematic search, *rp* is the relation category-index defined in the registration form, *q* is the query inserted by the user, *d* is a document to calculate the relevance, *SCD (q, d)* in (0,1), *RCD (q)* in (0,1), *RPUD* ≤ 1 and α + β + γ = 1. Notice that R (d) ≤ 1, being 1 the maximum relevance value for a document.

The parameters used in (1) are the following:

1. *SCD:* Similarity between the user's query *q* and the document *d*. To calculate this value, it is proposed to use the cosine formula considering in the Vector Space Model.
2. *RCD:* A matching function between the user's query *q* and the document *d,* expressing the relevance with regard to the relation category-index defined in the thematic search.
3. *RPUD:* A matching function between the user's query *q* and the document *d,* expressing the relevance with regard to the relation category-index defined in the registration interface.

In case of table 2, we are facing a user who focuses his search preferences on topics related to the environment with 60% percent prevalence, about culture 20% and politics about 20%. With this information, the IRSs must be able to enhance in the calculation of relevance the documents belonging to these categories, prioritizing those related to the environmental category. Another parameter that the algorithm uses is the categories (DC) to which each document stored belongs. To execute this task, an automatic categorization mechanism must be guaranteed that allows, once the documents are tracked and indexed, to assign in the data structure that represents each document in the collection, the category field with the calculated value. To solve the problem of a document belonging to more than one category, it is decided to proceed as with the user's profile and create an arrangement with the proportion of predominance of each category to which the document belongs, see an example in table 3.

Table 3 - Assignment of categories to a particular document (DC)

Category	Sports	Environment	Culture	Politics	Sciences
Index	0,1	0,2	0,6	0,05	0,05

The two scenarios which could be considered are the following:

Scenario 1. Users need objective results wich do not depend on the profile defined in the registration interface; in this case the value of β = 0,5, α = 0,5 and γ = 0. In this scenario, it is assumed that the most relevant documents are those that have a greater degree of similarity with the query, when applying the cosine formula and also that they belong to the categories of greater predominance in the thematic search.

Scenario 2. Users need results that vary only taking into their preferences selected in the registration form, in this case β = 0, γ = 0,5 and α = 0,5. In this scenario, it is assumed that the most relevant documents are those that have a greater degree of similarity with the query when applying the cosine formula and also that they belong to the highest scored categories in the user's search profile defined in the registration form interface.

These two scenarios are selected in automatic way by the IRS, if the user select the thematic search, then scenario 1 is selected in other case scenario 2 is selected. The relevance calculation for every document is defined by the execution of a series of steps shown in two procedures (rule 1 and rule 2) which are exemplified below. In case the results are framed in scenario 1 described above, then the USP value is replaced by the QC in the rules, if the results respond to scenario 2 then the value of the USP is used instead QC:

Rule 1

Input:

USP – Set of (*category*, *value*) of the user profile.

DC – Set of (*category*, *value*) associated to a document.

Output: RPUD

1: $I = \{c \mid (c,*) \in USP, (c,*) \in DC\}$ Ø

2: If $I == \emptyset$ then *RPUD* = 0

3: else

4: $(A,B) = \arg\max_{(a,b)} \{b \mid (a,b) \in USP, a \in I\}$

5: $(A,C) \in DC$

6: If B > C then RPUD = C // first case study

7: else RPUD = B

// second case study

Two case studies are designed to test the operation of the algorithm. In the first case study the percentage index of the categories of the user profile environment (0,6), politics (0,2), culture (0,2) and the categories of the document sport (0,1), environment (0,2), culture (0,7) are defined. For the first case rule 1 applies and the following values are defined: I = {Environment, Culture}, A = Environment, B = 0,6, C = 0,2, B > C → RPUD = 0,2. In the second case study, the percentage index of the categories of the user profile environment (0,3), politics (0,6), culture (0,1) and the categories of the document sport (0,1), environment (0,9) are defined. For the second case, rule 1 applies and the following values are defined: I = {Environment}, A = Environment, B = 0,3, **C**= 0,9, B < C → RPUD = 0,3 (rule 2, table 4).

Rule 2

Input:

USP – Set of (category, value) of the user profile.

DC – Set of (category, value) associated to a document.

Output: RPUD.

1: $I = \{c \mid (c,*) \in USP, (c,*) \in DC\}$

2: if $I == \emptyset$ then *RPUD* = 0

3: else

// Example 3

4: $AB = \{(a,b) \in USP \mid a \in I, b = \max\{v \mid (*,v) \in USP\}\}$

5: $(*,B) \in AB$

6: $C = \max\{c \mid (a,c) \in DC, (a,*) \in AB\}$

7: if $B > C$ then *RPUD* = *C*

8: else RPUD = B

Table 4 - Example 3: categories of the user's search profile and that of the documents

User's search profile	**Category**	**Environment**	**Politics**	**Culture**
	Percent index	0,45	0,45	0,1
Documents	Category	Environment	Politics	-
	Percent index	0,9	0,1	

In example 3, considering the USP and the DC of table 4 for a particular document, and applying rule 2, the values defined are the following: I = {environment, politics}, A = {environment, politics}, B = 0,45, D = {0,9, 0,1}, C = 0,9, B < C → RPUD = 0,45.

An experiment was carried out to evaluate the precision and recall metrics obtained before and after applying the algorithm in a Real Information Retrieval System (called Orión, developed by de University of Informatic Sciences) which uses as algorithm for the calculation of relevance a hybrid between the Probabilistic Model and the Vector Model based on the formula of cosine. Also, a second version of Orión was implemented according to expression (1). The experiments were conducted using a population of 23 users who have more than 5 years of experience in the use of information retrieval systems. In addition, 100 documents were selected, categorized according to the 6 categories defined in the Orión search engine. Each user selected a query of 3 proposals; experts, in IR and user profile modeling, annotated the documents that are considered relevant in relation to the selected query and the user's USP. The USP was obtained from the registration form as a relation category-index defined by users.

Results

The calculation of the relevance was executed for the two scenarios defined in the research but, due to lack of space in this paper, only scenario 2 of (1) will be addressed, using β = 0, α = 0,5 and γ = 0,5; thus *SCD* (*q*, *d*) and $RPUD(q,d,rp)$ functions are equally valued. Other scenarios and combinations of values of α, β and γ should be analyzed in further works. The results obtained can be seen in table 5.

Table 5 - Experimental results when applying the algorithm to calculate the relevance of the information

Average precision before applying the algorithm	Average precision after applying the algorithm	Average recall before applying the algorithm	Average recall after applying the algorithm
0,34	0,86	0,30	0,70

In both metrics there is a significant difference in the values obtained before and after applying the proposed algorithm, improving the quality of the results provided to users who interact with this search engine. In addition, these results shown that the preferences of the users and the categories of the stored documents play an important role in the calculation of the relevance of the documents returned in response to the questions of the users.

Conclusions

The analysis of the literature allowed to identify that there are inadequacies in the use of user preferences and document classifications to provide relevant and customized search results on information retrieval systems.

The design, scientific substantiation and implementation of the proposed algorithm integrates user profile preferences and document categories to retrieve relevant information.

The results obtained from applying the precision and recall metrics to the proposed algorithm allowed to corroborate that it improves the quality of the search results provided to users.

Bibliographic references

1. Marcos MC. Entrevista a *Ricardo Baeza-Yates*, de Yahoo! Investigation. Hipertext.net. 2008 [acceso: 16/03/2020];6:[aprox. 4 p.]. Disponible en: https://ddd.uab.cat/pub/artpub/2007/88758/hipertext_a2007n5a7/recuperacion-informacion.html

2. Searchenginejournal.com. Newtown Turnpike. EE. UU: Searchenginejournal.com; 2020 [acceso: 16/03/2020]. Disponible en: https://www.searchenginejournal.com/google-confirms-maccabees-algorithm-update/228901/

3. Gonzalo C, Codina L, Rovira, C. Recuperación de Información centrada en el usuario y SEO: categorización y determinación de las intenciones de búsqueda en la Web. Ind Comunic. 2015;5(3):19-27.

4. Sust E, Cuevas A, José O. Análisis de tendencias en la personalización de los resultados en buscadores web. RCCI. 2018;12(2):111-28.

5. Babekr STF, Khaled M. Personalized semantic retrieval and summarization of web based documents. Internat J Adv Comp Sc App. 2013;4(1):177-86.

6. Bibi T, Dixit P. Web search personalization using machine learning techniques. In: IEEE International Advance Computing Conference (IACC). IEEE; 2014. p. 1296-9.

7. Bostan S, Ghasemzadeh G. Personalization of Search Engines, Based-on Comparative Analysis of User Behavior. J Advan Computer Res. 2015;6(2):65-72.

8. Dumais ST. Personalized Search: Potential and Pitfalls. CIKM; 2016. p. 689.

9. Gao Q, Young I. A multi-agent personalized ontology profile based query refinement approach for information retrieval: control, automation and systems (ICCAS). 13th International Conference on IEEE. p. 2013:543-7.

10. Ghorab MR. Personalised Information Retrieval: survey and classification. Springer. 2013;23(4):381-443.

11. Hannak A. Measuring personalization of web search. In: Proceedings of the XXII International Conference on World Wide Web. ACM; 2013. p. 527-38.

12. Johnson MS. Personalized Recommendation System for Custom Google Search. International Journal of Computer & Mathematical Sciences; 2016:5.

13. Fransson J. Efficient Information Searching on the Web. A Handbook in the Art of Searching for Information; 2010.

14. Searchengineland.com. Newtown Turnpike. EE.UU.: Searchengineland.com; c2020 [acceso: 16/03/2020]. 8 major Google algorithm updates, explained [aprox. 15 p.]. Disponible en: https://searchengineland.com/8-major-google-algorithm-updates-explained-282627

15. Baquerizo R, Leyva P, Febles J, Viltres H, Estrada V. Algorithm for calculating relevance of documents in information retrieval systems. IRJET; 2017 [acceso: 16/03/2020];4(3):[aprox. 13 p.]. Disponible en: https://www.irjet.net/archives/V4/i5/IRJET-V4I501.pdf

16. Ortega P, Leyva P, Febles JP, Viltres H, Delgado Y. Computational model for the processing of documents and support to the decision making in systems of information retrieval. Internat Res J Engin Technol. IRJET; 2017 [acceso: 16/03/2020];4(5): [aprox. 17 p.]. Disponible en: https://www.irjet.net/archives/V4/i5/IRJET-V4I502.pdf

17. Viltres H, Rodríguez P, Febles JP, Estrada V. Information retrieval with semantic annotation. Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology; 2019 [acceso: 16/03/2020];[aprox. 10 p.]. Disponible en: http://laccei.org/LACCEI2019MontegoBay/full_papers/FP308.pdf

Conflict of interest

The authors declare that we have no conflict of interest in the development of this research.

Authorship statement

All the signing authors were involved in the elaboration of the theoretical framework, development and validation of the algorithm.