Revista Facultad de Ingeniería
ISSN: 0121-1129
ISSN: 2357-5328
Universidad Pedagógica y Tecnológica de Colombia

Viltres-Sala, Hubert; Estrada-Sentí, Vivian; Febles-Rodríguez, Juan-Pedro; Jiménez-Moya, Gerdys-Ernesto

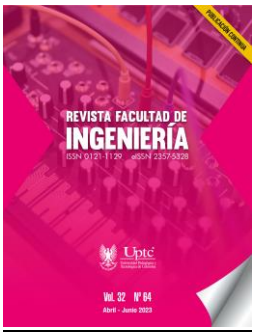
Information Retrieval Model with Query Expansion and User Preference Profile

Revista Facultad de Ingeniería, vol. 32, no. 64, 4, 2023, April-June
Universidad Pedagógica y Tecnológica de Colombia

DOI: https://doi.org/10.19053/01211129.v32.n64.2023.15208

Available in: https://www.redalyc.org/articulo.oa?id=413975585004

Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Information Retrieval Model with Query Expansion and User Preference Profile

Hubert Viltres-Sala[1]
Vivian Estrada-Sentí[2]
Juan-Pedro Febles-Rodríguez[3]
Gerdys-Ernesto Jiménez-Moya[4]

Received: February 02, 2023 **Accepted:** May 01, 2023 **Published:** May 09, 2023

Citation: H. Viltres-Sala, V. Estrada-Sentí, J.-P. Febles-Rodríguez, G.-E. Jiménez-Moya, "Information Retrieval Model with Query Expansion and User Preference Profile," *Revista Facultad de Ingeniería*, vol. 32, no. 64, e15208 2023. https://doi.org/10.19053/01211129.v32.n64.2023.15208

Abstract

Understanding the user's search intention enables identifying and extracting the most relevant and personalized search results from the available information, according to the user's needs. This paper proposes an algorithm for relevant

[1] M. Sc. Universidad de Ciencias informáticas (Habana, Cuba). hviltres@uci.cu. ORCID: 0000-0002-5116-3665
[2] Ph. D. Universidad de Ciencias informáticas (Habana, Cuba). vivian@uci.cu. ORCID: 0000-0002-7513-7891
[3] Ph. D. Universidad de Ciencias informáticas (Habana, Cuba). febles@uci.cu. ORCID: 0000-0003-3126-7667
[4] Ph. D. Universidad de Ciencias informáticas (Habana, Cuba). gejimenez@uci.cu. ORCID: 0000-0002-0146-4953

information retrieval that combines user preferences profile and query expansion to get relevant and personalized search results. The information retrieval process is validated using Precision, Recall and Mean Average Precision (MAP) metrics applied to a dataset that contains the standardized documents and preferences profiles. The results allowed us to demonstrate that the algorithm improves the information retrieval process by finding documents with better quality and greater relevance to the users' needs.

Keywords: personalized information retrieval; query expansion; semantic annotation; user profile.

Modelo para la recuperación de información con expansión de consulta y perfil de preferencia de los usuarios

Resumen

Comprender la intención de búsqueda del usuario permite identificar y extraer los resultados de búsqueda más relevantes y personalizados de la información disponible según sus necesidades. En el presente artículo se plantea un algoritmo para la recuperación de información relevante que combina las preferencias del perfil del usuario y la expansión de consulta para obtener resultados de búsqueda relevantes y personalizados. El proceso de recuperación de información se valida mediante las métricas de *Precision, Recall* y *Mean Average Precision (MAP*) aplicadas a un conjunto de datos que contiene los documentos estandarizados y los perfiles de preferencias. Los resultados permitieron demostrar que el algoritmo mejora el proceso de recuperación de información al arrojar documentos con mejor calidad y relevancia según las necesidades de los usuarios.

Palabras clave: anotación semántica; expansión de consulta; perfil de usuario; recuperación de información personalizada.

Modelo de recuperação de informações com expansão de consulta e perfil de preferência do usuário

Resumo

Compreender a intenção de pesquisa do usuário permite identificar e extrair os resultados de pesquisa mais relevantes e personalizados das informações disponíveis de acordo com suas necessidades. Neste artigo, é proposto um algoritmo para a recuperação de informações relevantes que combina as preferências do perfil do usuário e a expansão da consulta para obter resultados de pesquisa relevantes e personalizados. O processo de recuperação da informação é validado por meio de métricas Precision, Recall e Mean Average Precision (MAP) aplicadas a um conjunto de dados contendo documentos padronizados e perfis de preferência. Os resultados permitiram demonstrar que o algoritmo aprimorou o processo de recuperação da informação ao produzir documentos com melhor qualidade e relevância de acordo com as necessidades dos usuários.

Palavras-chave: anotação semântica; expansão da consulta; perfil de usuário; recuperação de informações personalizadas.

I. INTRODUCTION

Traditional search engines generally use statistical techniques that determine relevance by keyword matching without fully understanding the user's search intentions and the implicit context of the indexed information [1-2]. Retrieving relevant and personalized information poses a challenge for Information Retrieval Systems (IRS) that need to satisfy a user's need presented in the form of a question and specified through a set of keywords by analyzing a collection of documents with variable volume and format [3]. The Information Retrieval (IR) process has been improved by developing several models [4-8] focused on solving problems related to:

- The difficulty to identify and understand the user's needs written in natural language when entering exact or ambiguous terms limits the retrieval of relevant documents.
- The retrieval of documents by the statistical term without analyzing the context of the question and the stored information.

To solve the main shortcomings identified in the IRSs, we propose to process documents [9-11], queries [12-15], and the user profile [2,16,17] to improve the IR by using ontologies and analyzing the user's behavior.

In the IR process —even though knowing users' search intent is important to understand their need— applying relevance algorithms that combine preference profile [16-18], query expansion [2,14], and semantic annotations [2,9,11] enables enhancing the obtained search results.

Customization of search results plays an important role in the level of user satisfaction when interacting with an IRS [19]. Gupta et al. [20] proposed the use of web mining techniques, Natural Language Processing, and ontologies to extract behavioral patterns from users' web logs. Regarding the generation of the user preferences profile, the main proposals focus on combining explicit (selected interests in their profile related to topics, age, and gender) and implicit preferences (browsing history and user location) to model the profile according to the users' behavior [2, 16,17,21,22].

According to [23,24] in the construction of a user profile, three fundamental phases related to collecting, constructing, and using the data acquired by analyzing the user's behavior in a CRS are identified. IRSs register the user's actions to collect as much information as possible to identify behavioral patterns and obtain their preferences.

Nandanwar, Choudhary, and Singh [25] analyze users' implicit and explicit preferences to generate a hybrid profile, and they weight it by a constant value $\alpha = 0.6$ between a short-term and long-term profile to obtain a preference value between [0, 1]. However, the proposal does not describe the method to apply the temporality of the profiles. Queries are classified into ambiguous and unambiguous to improve their processing; the terms of the ambiguous query are expanded with WordNet and modified with the combination of the maximum similar category and the user's profile. The authors apply cosine similarity to determine the relevance between the expanded query and the document collection to display results relevant to the user and then update the profile preferences [25].

In addition, a user preference profile combining tags, annotations, and retrieved documents [18]. Two algorithms that extract the terms are designed to analyze their match with the documents and select the most relevant ones to be included in the user profile. The user's preference profile is used to expand the query.

The proposals by [23,24,25] focus on extracting information from user behavior to predict search preferences. The preferences elicitation and the temporality of the profile present shortcomings, which reduce the effectiveness of the search results customization. This paper proposes weighing the preferences profile temporality and consider the relevance of the concepts used in the query and documents retrieved and consulted by the user to fix the identified deficiencies.

Query expansion (QE) seeks to reformulate the original words written by the user and substitute them for relevant terms from different sources, weighing them to reduce ambiguities and improve the information retrieval accuracy [2,13,14]. When the user submits a query, it is expressed in natural language and sometimes uses imprecise sentences that do not allow the IRSs to provide relevant results. Some literature reports [12,14,26] propose that, to understand the context of the question

asked by the user, it is necessary to expand the query through ontological repositories and preference profiles [13,14].

The main expansion techniques in [2,12,13,14] focus on processing the entered terms through a similarity measure with an ontological repository to obtain semantically similar concepts included in the query reformulation. To reduce ambiguities and personalize the results, it has been proposed to include information related to the user's preference profile in the expansion process [2,12,13,17].

The use of information annotation in QE is evidenced in [2], which suggests combining information from clinical diagnoses with word embedding to retrieve relevant biomedical information. Besides, three types of words are included (domain-specific, domain-related, and hybrid) to perform the annotations in the original query. Also, MetaMap is used to recognize and extract biomedical concepts that are candidates to be included in the query expansion. In the selection of candidate terms, the cosine similarity is used, and the original query is weighed with the expanded terms.

The method proposed by Nandanwar, Choudhary, and Singh [25] combines the entered query with the user profile and WordNet to customize the expansion. The queries are initially classified into ambiguous or unambiguous. WordNet and the most similar category in the user's profile are used for ambiguous queries. For unambiguous queries, all the terms of the query are identified in WordNet, and the most representative term in the user's profile is added to the original query.

Suma [5] proposes to expand the query with user profile information and a domain ontology by weighing semantic relations to select candidate terms for expansion. Also, Jain et al. [14] proposed a method for the construction of an ontology to identify semantic relationships in query expansion. In addition, Dahir et al. [13] employed a query expansion method based on bag-of-words distribution to select the 20 terms with the highest similarity and determine the efficient expansion terms from single or multi-valued DBpedia attributes.

Analyzing the consulted solutions allows identifying the variety of similarity measures to recognize the candidate concepts to be included in the expansion. In addition, the main shortcomings are related to the use of data sources and the amount of

information to be added. Our research proposes to expand the terms using an ontological repository and to reduce ambiguity through the user's preference profile. As a novelty, concepts from the user's profile with greater similarity to the query are included.

Relevance is defined in [27,28] as the level of usefulness, quality, and value of the retrieved information to satisfy the user's needs. Relevance can be represented by a scalar and measures the degree of agreement of different criteria (match, meaning, and frequency of occurrence of the terms) between the entered query and the document collection.

Wan et al. [29] presented a method that combines similarity by matching and semantics to obtain an initial score of the documents using the BM25 algorithm. Then, they improved the definition of relevance with the linear combination of the semantic analysis by BERT and reformulated the query using the terms extracted from the N top-ranked documents to improve the search results.

A hybrid model for document ranking that combines query probability and semantic similarity between concepts using WordNet was proposed by Neji et al. [30]. This model calculates the similarity between concepts by Jiang-Conrath similarity and user preferences to extract concepts and reformulate the query. It determines the similarity between documents and the query to sort the search results.

The analysis of previous works allowed us to identify a coincidence in the use of query expansion [12-15], user profile processing [2,16,17], and documents [9,10,13] to retrieve relevant and customized information.

This article is organized as follows: Section 1 presents a selection of works related to the user's profile, the query expansion, and customized information retrieval; Section 2 presents the proposed model and describes the implemented algorithms; Section 3 shows the discussion based on the results obtained in this research; and Section 4 presents the conclusions.

II. Materials and Methods

This section describes the proposed model that enables collecting, indexing, and processing the information available on the Web to obtain relevant and personalized

search results. The main objective of the model is to improve the quality of the search results provided to the users when they access an IRS by integrating the preference profile, query expansion, and stored documents into a relevance algorithm (Figure 1).

When the users register in the IRS, they create their profile, and the system analyzes their behavior and uses this information to personalize the search results. The use of an ontological repository to make semantic annotations and expand the query makes it possible to improve information processing and retrieve the most relevant documents for the user.

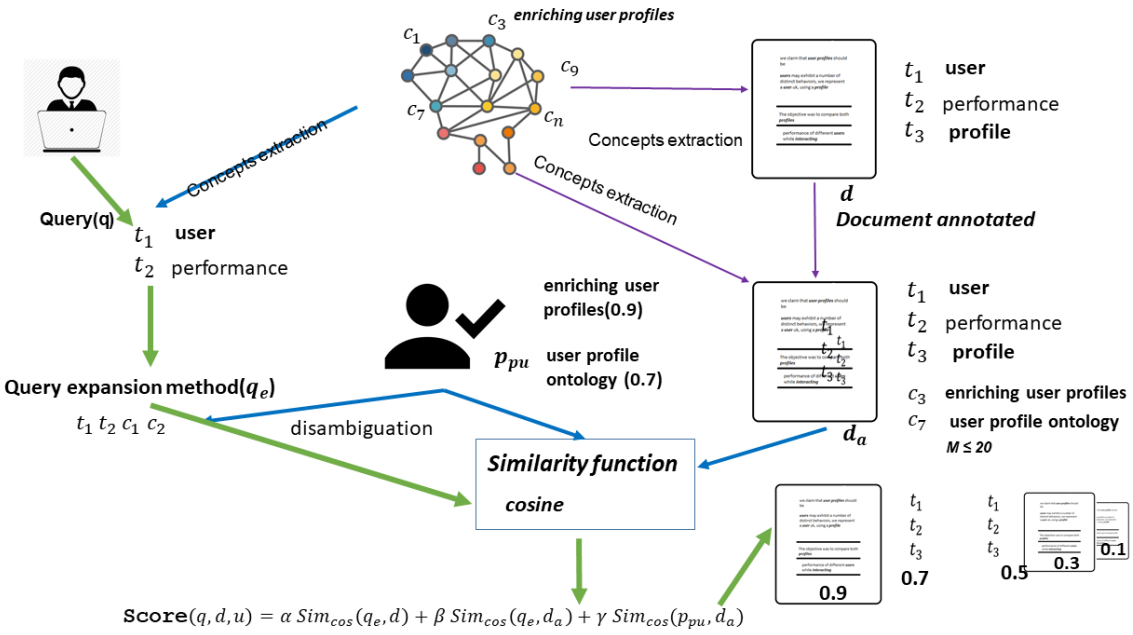


Fig. 1. Retrieval of information based on query expansion and user preference profile.

The model obtains the query entered by the user as inputs and, through the indexing process, the documents extracted from the Web (Figure 1). The documents are categorized and semantically annotated to enrich and reduce their semantic ambiguity by means of an ontological repository. Besides, the expansion process is applied to the entered query, and the relevance algorithm is used to retrieve the documents that best fit the user's needs. The outputs are the search results relevant to the user and the user's updated preference profile.

A. Query Expansion with Ontology and User Preference

Query expansion reformulates the original query by incorporating relevant terms from different sources, weighting them in the expansion process to understand their semantic meaning and improve the information retrieval accuracy [2]. Query processing mainly aims at disambiguating the terms entered by the user to improve the relevance of the search results.

The developed expansion algorithm (Algorithm 1) combines the original query terms, a domain ontology, and the user's preference profile. The entered terms are processed to obtain the candidates to be included in the expansion. The similarity measure between concepts, proposed by Rafa et al. [2] (Expression 1), is applied to obtain the most relevant concept to be included in the expanded query.

Algorithm No. 1. Query Expansion.

Input: Set of terms *t* of the query, *q*; user profile preferences, *ppu;* and ontological repository, *o*.
Output: expanded query, *qe*.

Begin algorithm

Input

O ontological repository with the structure of concepts c(c1, c2, c3,..., cn) and relations *r*

PCAu={pu1,pu2,pu3…pun} user preference profile **u**

t={t1,t2,t3…tn} ***€ q*** query term set **q**

CA: most relevant concept of the profile not included in the expansion

CCi: candidate concepts to expand

WCCi: weight of the candidate concept

Ce: expanded concept

Cpuk: concept associated with the user profile

WCEi: concept weight Ce

Output

qe={t1,c2,c3} expanded query

1. Get for $\{\boldsymbol{t_1}, \boldsymbol{t_2}, \ldots \boldsymbol{t_n}\} \in \boldsymbol{q}$ los *cc* in ontology and WordNet
2. Get ***q*** and ***ppu***

   Requiere *o* ← ontological repository

   Requiere *ppu* ← get preference profile

for i← **1** *to i* ← $\boldsymbol{n}$ *do*

Calculate *CCi*= SimLin(ti,CCj) CCi ← get cc for the term ti ∈q

CCi ← *Get (CCi, SimLin(ti,CCj)*

for k← **1** *to k* ← $\boldsymbol{m}$ *do*

Simcos(CCi,Cpuk)

End for

if Simcos(CCi,Cpuk)=> α then

WCEi = (WCCi + Cpuk) /2

else

WCEi = WCCi/2

qe ←Add(WCEi)

end if;

end;

end;

3. Get most similar concept

for k← **1** *to k* ← $\boldsymbol{m}$ *do*

CA ← Get(CCi, Simcos(WCEi,CCk)

End

Reformulate *qe* ← *qe* ∩ *CA*

Output qe

End algorithm

The query disambiguation process is performed according to the user's preferences profile to obtain concepts associated with the entered terms. As a first step of the algorithm (1), the user's search history (PCA) PCAu={pu1,pu2,pu3...pun} is obtained from the user's profile to access the relevance of the concepts stored in his preferences profile. Then, the original query terms are retrieved, and the similarity between the candidate and associated concepts is calculated to determine the relevance of the information retrieved from the user's profile. As a second step, the candidate concepts for the expansion are obtained, and the Cosine similarity is applied to select the most relevant concepts to reformulate the query. The result is an expanded query disambiguated by user profile and ontology.

B. Relevant and Customized Information Retrieval

Algorithm #2 is developed and applied to determine the similarity between the query and the stored documents, which combines query expansion and document annotation with a result weighing according to the user search preferences. For each document, there is a representation of the original document vector (term/weight) and annotated document vector (concepts/index).

Algorithm #2 uses the documents (d), the user's preference profile (ppu), and the expanded query (eq) to determine the relevance of the search results. The relevance of a document according to the query entered by the user is calculated using Expression (1), and the relevance calculation algorithm is designed and implemented:

$$\text{Score}(q,d,u) = \alpha\, Sim_{cosine}(q,d) + \beta\, Sim_{cosine}(q_e,d_a) + \gamma\, Sim_{cosine}(p_{pu},d_a) \qquad (1)$$

Where Score (q,d,u) is defined as the relevance for the user (u) of document (d) according to the similarity between the query (q), the document (d), and the preference profile (ppu). It is assumed that the results obtained when executing the same query by several users do not necessarily satisfy them to the same extent. In the developed algorithm, a linear combination is defined to weigh values α, β, γ, where the value of Score (q,d,u) <=1 provided that α + β+ γ=1. Also, the Cosine measure (2) was used, and it obtains values in the interval [0,1] to determine the similarity. Besides, the Cosine similarity (Sim_{cosine}) expression establishes the significance.

$$Sim_{cosine}(\mathrm{d},\mathrm{q}) = \frac{\sum \mathrm{W_{t,d}\ x\ W_{t,q}}}{\sqrt{\sum_t \mathrm{W^2_{t,d}}}\ \mathrm{x}\ \sqrt{\sum_t \mathrm{W^2_{t,q}}}} \qquad (2)$$

In addition, it is stated that:

- Cosine similarity (q,d): function to calculate the similarity between the query (qe) and the document (d).
- Cosine similarity (qe,da): function to calculate the similarity between the query (qe) and the annotated document (da).
- Cosine similarity (ppu,da): function to calculate the similarity between the concepts annotated in the documents (da) and the user's search concept preference profile (ppu).

Algorithm No. 2. Calculation of relevance.

Input: Set of annotated documents d from collection, D; user profile preferences, ppu; and query, q.

Output: relevant documents.

Begin algorithm

Input

{d1,d2,d3…dn} € ***D*** Documents from Collection **D**

Ppu={p1,p2,p3…pn} user preferences profile u

t={t1,t2,t3…tn} € **q** query term set **q**

Output

$$d = \begin{Bmatrix} d1_1, d2_1, d7_{0.9}, d9_{0.7}, d109_{0.6}, \\ d22_{0.2}, dn_n \end{Bmatrix}$$

1. Get *qe* and *ppu*

   Requiere *qe* ← expand query q whith algorithm 1

   Requiere ppu ← get preference profile

2. **For** $\{\boldsymbol{d_1}, \boldsymbol{d_2}, \boldsymbol{d_3}, \ldots \boldsymbol{d_n}\} \in \boldsymbol{D}$ ***for*** *i*← **1** *to i* ← $\boldsymbol{n}$ *do*

Calculate $Score(q,d,u) = \alpha\ Sim_{cos}(q,d) + \beta\ Sim_{cos}(q_e,d_a) + \gamma\ Sim_{cos}(p_{pu},d_a)$ (3)

   End for

3. Return relevant documents sorted by $Score(q,d,u)$

End algorithm

III. RESULTS

To evaluate the proposed model, an experiment was designed to measure the relevance of retrieved documents by integrating user preferences, query expansion, and relevance calculation. The experiment to test the search results quality is described below.

From the collection of stored documents, 5000 documents were selected, categorized, and semantically annotated using domain ontologies and WordNet. Each document contains the original and the concepts annotated by semantic similarity with the ontology repository. Thirty user profiles were designed. The preferences of each profile were determined by recording 10 queries with 20 relevant documents and an average of 70 annotated concepts in its history. In addition, implicit data related to their preferences, search categories, location, and other data recorded by the IRS are captured. Each user was assigned 50 queries with different

levels of semantic ambiguity (high, medium, and low) and executed in the 4 scenarios designed to perform the search. Each user was informed of the characteristics of the assigned profile and was asked to simulate the defined preferences to evaluate the search results.

For the selection of documents, queries, user profiles, and scenarios, the studies conducted by Hahm et al. [32], Xu et al. [33], and Malik et al. [15] —who designed similar proposals and used metrics such as Precision (P), Recall, and Mean Average Precision (MAP) for validation— were considered. Due to the importance of having a correct data source to select and collect documents, queries, and user profiles, the following criteria were considered:

- Quality: documents should be high-quality and semantically annotated to ensure the relevance evaluation of the developed system. Queries should be expanded using an ontology to improve their relevance and performance. User profiles should contain detailed information about preferences and relevant concepts.
- Relevance: documents should be relevant to the application domain and contain a wide variety of topics. Queries should be varied and represent different levels of complexity. User profiles should be as similar as possible to the user's preferences and needs in the application domain.
- Diversity: documents, queries, and user profiles should be varied to ensure that the system performs well in different scenarios and situations.
- Representativeness: the selected documents, queries, and user profiles should accurately represent the application domain and user interests.

In relation to the execution of the data, a case study composed of 4 scenarios was designed to compare the functioning of the IRS. These scenarios cover the operation of the IRS without modifications, with query expansion, with user preferences, and integrating the preferences' relevance and query expansion.

- Base proposal (without modification (PS)): measures system performance without modification and is set as execution condition for $\alpha = 1$, $\beta = 0$, and $\gamma = 0$ for Expression 1.

- Proposal # 2 with query expansion (Pqe): measures the performance of the system with query expansion and is set as execution condition for α = 0, β = 1 and γ = 0 for Expression 1.
- Proposal #3 with query expansion and customization by user search history preferences (Pqeu): measures the performance of the system with query expansion and weighing by search history. It is set as an execution condition for α = 0, β = 0.5, and γ = 0.5 for Expression 1.
- Proposal #4 with terms, query expansion, and search preferences (Pdqeu): measures system performance with keyword search, query expansion, and search history weighing. It is set as an execution condition for α = 0.3, β = 0.5, and γ = 0.2 for Expression 1.

The selection of weighing values for α, β, and γ is based on the importance of semantic annotation for information retrieval. The more relevant the semantic annotation, the more similar the query, the user profile, and the indexed documents. Query expansion increases the coverage of possible relevant terms and generates higher weighing values compared to terms and the user profile.

The expanded query has a higher probability of including relevant terms that may be present in relevant documents but were not included in the original query. Terms and user profile have more limited coverage and are less likely to be present in all relevant documents because they are more specific and limited to the keywords provided by the user in the original query and their preference profile.

Next, the queries assigned to each user profile are executed, and the results are recorded to apply the Precision (P), Recall (R), and Mean Average Precision (MAP) metrics to measure the quality of the results.

Precision: retrieved documents relevant to the user's information need, set as cut-off measure P@10 and determined by Expression 4.

$$\mathrm{P@10} = \frac{\text{No of Relevant Documents in the top 10 documents}}{10} \qquad (4)$$

Completeness: the documents relevant to a query were retrieved and determined by Expression 5.

$$\mathrm{R} = \frac{\text{Relevant documents retrieved}}{\text{Relevant Documents}} \qquad (5)$$

Mean Precision is the average of the mean precision scores for each query and is determined by Expression 6.

$$\boldsymbol{MAP} = \frac{\sum_{q=1}^{Q} \boldsymbol{AveP}_{(q)}}{\boldsymbol{Q}} \qquad (6)$$

Table 1. P, R, MAP metrics results.

Approach	P@10	R@10	MAP@10
PS	0.29	0.19	0.17
Pqe	0.55	0.43	0.36
Pqeu	0.63	0.54	0.45
Pdqeu	0.86	0.83	0.71

Expression 4 was used to calculate the precision of the results, and P@10 was selected as the cut-off ranking. The use of a ranking to evaluate the first results is common in research [2,13,31].

The analysis of the results obtained after executing the case study confirmed that combining user preferences and ontologies with semantic annotation, query expansion, and relevance calculation improves the IR. The precision values obtained for P@10 (0.86) were acceptable and allowed to corroborate that query expansion improves search results by 2.3 times the system without modification and increases up to 3 times when combined with document annotation and user preferences profile. Better results are obtained when the documents are annotated, and the user profile allows them to retrieve enough information to define their preferences.

To compare the results obtained with research analyzed in the related works section, versions of the proposals of [2,13,31] were implemented according to the descriptions because they address query expansion and user profiles. A new experiment was designed with 10 users, 15 queries, and 1000 documents to compare the implemented proposals.

It is assumed that the most relevant documents have the highest degree of similarity with the expanded query when applying the cosine measure and that the concepts annotated in the document and the query have the highest similarity with the preferences defined in the user's profile. For Experiment 2, $\alpha = 0.3$, $\beta = 0.5$, and $\gamma = 0.2$ are defined as the execution condition for weighing the values in Expression 4.

Precision (P), Recall, and Mean Average Precision (MAP) metrics are used to evaluate the results, and they are shown in Table 2.

Table 2. P, R, MAP metrics results. Experiment 2.

Approach	P@10	R@10	MAP@10
Hahm et al. [32]	0.64	0.61	0.60
Xu et al. [33]	0.73	0.67	0.46
Malik et al. [15]	0.89	0.65	0.62
Proposed model	0.87	0.83	0.70

The values of P@10 and R@10 were higher than those of research [13,31], similar to [2] and evidence, and had greater relevance when using information annotation, user profiles, and queries. Also, the results obtained from MAP (0.70) evidence similarity with proposals [2,13,31], which include search customization with user preference profiles and query expansion. Better results are evidenced when the user profile is well defined, thus helping to decrease the ambiguity of the terms entered in the query.

The evaluation using the Precision (P), Recall (R), and Mean Average Precision (MAP) metrics demonstrates the quality, relevance, and pertinence of information retrieval with semantic annotation. It also reveals that the integration of user preferences and semantic annotation improves query expansion and adjusts the search results to their needs by selecting the most relevant ones to increase their experience in using IRS.

IV. Discussion and Conclusions

The analysis of the information retrieval process allowed identifying the main deficiencies that affect information processing, e.g., information overload, heterogeneity of information sources, and interoperability, which greatly hinder the adequate processing of the available information. The developed query expansion algorithm combines user preferences and the concepts associated with the entered terms to help reduce ambiguity and retrieve relevant and customized documents. Also, the semantic annotation of information in the documents, the user profile, and the query expansion increase the relevance of the retrieved documents.

The application of the relevance algorithm improves the process of retrieving customized and relevant information. Applying and analyzing Precision, Recall, and Mean Average Precision (MAP) metrics to the developed proposal allows us to verify that it improves the quality of the results.

ACKNOWLEDGEMENTS

Thanks to the Erasmus+ KA107 International Dimension (Associated Countries) program of the University of Granada, the Department of Computer Science and Artificial Intelligence, the Center for Research in Information and Communication Technologies of the University of Granada (CITIC-UGR), and the University of Computer Science for their support in the development of the research.

AUTHORS' CONTRIBUTION

Hubert Viltres-Sala: Investigation, Formal Analysis, Methodology, Writing-original draft.

Vivian Estrada-Sentí: Supervision, Methodology, Validation

Juan-Pedro Febles-Rodríguez: Supervision, Methodology, Validation.

Gerdys-Ernesto Jiménez-Moya: Supervision, Methodology, Validation, Writing-Review and Editing.

References

[1] H. Viltres, P. Leyva, J. P. Febles, V. Sentí, "Information retrieval with semantic annotation," in *17th LACCEI International Multi-Conference for Engineering, Education, and Technology*, 2019. https://doi.org/10.18687/LACCEI2019.1.1.308

[2] T. Rafa, S. Kechid, "Semantic Representation of a Geo-Social User Profile for a Personalised Information Retrieval," *Journal of Information and Knowledge Management*, vol. 20, no. 4, e2150044, 2021. https://doi.org/10.1142/S0219649221500441

[3] P. P. Joby, "Expedient information retrieval system for web pages using the natural language modeling," *Journal of Artificial Intelligence*, vol. 2, no. 2, pp. 100-110, 2020. https://doi.org/10.36548/jaicn.2020.2.003

[4] S. Sengan, G. K. Kamalam, J. Vellingiri, J. Gopal, P. Velayutham, V. Subramaniyaswamy, "Medical information retrieval systems for e-Health care records using fuzzy based machine learning model," *Microprocessors and Microsystems*, In-Press, e103344, 2020. https://doi.org/10.1016/j.micpro.2020.103344

[5] V. Suma, “A novel information retrieval system for distributed cloud using hybrid deep fuzzy hashing algorithm,” *JITDW*, vol. 2, no. 3, pp. 151-160, 2020. https://doi.org/10.36548/jitdw.2020.3.003

[6] A. Jalilifard, V.F. Caridá, A. F. Mansano, R. S. Cristo, F. P. C. da Fonseca, “Semantic sensitive TF-IDF to determine word relevance in documents,” in *Advances in Computing and Network Communications*, 2021, pp. 327-337. https://doi.org/10.1007/978-981-33-6987-0_27

[7] S. Zhuang, H. LI, G. Zuccon, “Deep query likelihood model for information retrieval,” in *European Conference on Information Retrieval*, 2021. pp. 463-470. https://doi.org/10.1007/978-3-030-72240-1_49

[8] X. Liao, Z. Zhao, “Unsupervised approaches for textual semantic annotation, a survey,” *ACM Computing Surveys*, vol 52, no. 4, pp. 1-45, 2019. https://doi.org/10.1145/3324473

[9] D. Di Caprio, F. J. Santos-Arteaga, M. Tavana, “An information retrieval benchmarking model of satisficing and impatient users' behavior in online search environments,” *Expert Systems with Applications*, vol. 191, e116352, 2022. https://doi.org/10.1016/j.eswa.2021.116352

[10] S. Albukhitan, A. Alnazer, T. Helmy, “Framework of semantic annotation of Arabic document using deep learning,” *Procedia Computer Science*, vol. 170, pp. 989-994, 2020. https://doi.org/10.1016/j.procs.2020.03.096

[11] W. Wei, Q. Wu, D. Chen, Y. Zhang, W. Liu, G. Duan, X. Luo, “Automatic image annotation based on an improved nearest neighbor technique with tag semantic extension model,” *Procedia Computer Science*, vol. 183, pp. 616-623, 2021. https://doi.org/10.1016/j.procs.2021.02.105

[12] H. K. Azad, A. Deepak, “Query expansion techniques for information retrieval: a survey,” *Information Processing and Management*, vol. 56, no. 5, pp. 1698-1735, 2019. https://doi.org/10.1016/j.ipm.2019.05.009

[13] S. Dahir, A. “El Qadi, A query expansion method based on topic modelling and DBpedia features,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, e100043, 2021. https://doi.org/10.1016/j.jjimei.2021.100043

[14] S. Jain, K. R. Seeja, R Jindal, “A fuzzy ontology framework in information retrieval using semantic query expansion,” *International Journal of Information Management Data Insights*, vol. 1, no. 1, e100009, 2021. https://doi.org/10.1016/j.jjimei.2021.100009

[15] S. Malik, U. Shoaib, S. A. C. Bukhari, H. El Sayed, M. A. Khan, “A hybrid query expansion framework for the optimal retrieval of the biomedical literature,” *Smart Health*, vol. 23, e100247, 2022. https://doi.org/10.1016/j.smhl.2021.100247

[16] S. Abri, R. Abri, S. Çetin, “Group-based personalization using topical user profile,” in *28th ACM Conference on User Modeling, Adaptation and Personalization*, 2020, pp. 181-186. https://doi.org/10.1145/3386392.3399559

[17] Z. Ma, Z. Dou, Y. Zhu, H. Zhong, J. R Wen, “One Chatbot Per Person: Creating Personalized Chatbots based on Implicit User Profiles,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 555-564. https://doi.org/10.1145/3404835.3462828

[18] D. Zhou, X. Wu, W. Zhao, S. Lawless, J. Liu, “Query expansion with enriched user profiles for personalized search utilizing folksonomy data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1536-1548, 2017. https://doi.org/10.1109/TKDE.2017.2668419

[19] M. Bravo, A. Aldea, L. F. Hoyos-Reyes, “Automated Ontology Population and Enrichment of Scientific Publications,” *Journal of Physics: Conference Series*, vol. 1, e012139, 2021. https://doi.org/10.1088/1742-6596/1828/1/0121

[20] K. Gupta, N. Sachdeva, V. Pudi, “Explicit modelling of the implicit short term user preferences for music recommendation,” in *European Conference on Information Retrieval*, 2018. pp. 333-344. https://doi.org/10.1007/978-3-319-76941-7_25

[21] J. Choudhary, D. S. Tomar, D. P. Singh, “An Efficient Hybrid User Profile Based Web Search Personalization Through Semantic Crawler,” *National Academy Science Letters*, vol. 42, no. 2, pp, 105-108, 2019. https://doi.org/10.1007/s40009-018-0686-2

[22] F. Zarrinkalam, H. Fani, E. Bagheri, “Extracting, Mining and Predicting Users' Interests from Social Networks,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1407-1408. https://doi.org/10.1145/3292500.3332279

[23] S. Gauch, M. Speretta, A. Chandramouli, A. Micarelli, “User profiles for personalized information access,” in *The adaptive Web*, 2007, pp. 54-89. https://doi.org/10.1007/978-3-540-72079-9_2

[24] E. Vicente-López, L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, “Use of textual and conceptual profiles for personalized retrieval of political documents,” *Knowledge-Based Systems*, vol. 112, pp. 127-141, 2016. https://doi.org/10.1016/j.knosys.2016.09.005

[25] A. K. Nandanwar, J. Choudhary, D. P. Singh, “Web search personalization based on the principle of the ant colony,” *Procedia Computer Science*, vol. 189, pp. 100-107, 2021. https://doi.org/10.1016/j.procs.2021.05.073

[26] F. T. da Silva, J. E. Maia, “Query Expansion in Text Information Retrieval with Local Context and Distributional Model,” *Journal of Digital Information Management*, vol. 17, no. 6, e313, 2019. https://10.6025/jdim/2019/17/6/313-320

[27] M. Pereira, E. Etemad, F. Paulovich, “Iterative learning to rank from explicit relevance feedback,” in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 698-705. https://doi.org/10.1145/3341105.3374002

[28] J. Serrano-Guerrero, F. P. Romero, J. A. Olivas, “A relevance and quality-based ranking algorithm applied to evidence-based medicine,” *Computer methods and programs in biomedicine*, vol. 191, e105415, 2020. https://doi.org/10.1016/j.cmpb.2020.105415

[29] J. Wang, M. Pan, T. He, X. Huang, X. Wang, X. Tu, “A Pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval,” *Information Processing and Management*, vol. 57, no. 6, e102342, 2020. https://doi.org/10.1016/j.ipm.2020.102342

[30] S. Neji, T. Chenaina, A. M. Shoeb, L. B. Ayed, “HIR: a hybrid IR ranking model,” in *IEEE 45th Annual Computers, Software, and Applications Conference*, 2021. pp. 1717-1722. https://10.1109/COMPSAC51774.2021.00256

[31] B. Selvalakshmi, M. Subramaniam, “Intelligent ontology based semantic information retrieval using feature selection and classification,” *Cluster Computing*, vol. 22, no. 5, pp. 12871-12881, 2019. https://doi.org/10.1007/s10586-018-1789-8

[32] G. J. Hahm, M. Y. Yi, J. H. Lee, H. W. Suh, “A personalized query expansion approach for engineering document retrieval,“ *Advanced Engineering Informatics*, vol. 28, no 4, pp. 344-359, 2014. https://doi.org/10.1016/j.aei.2014.04.002

[33] B. Xu, H. Lin, L. Yang, K. Xu, Y. Zhang, D. Zhang, Z. Yang, J. Wang, Y. Lin, F. Yin, “A supervised term ranking model for diversity enhanced biomedical information retrieval,” *BMC bioinformatics*, vol. 20, no 16, e590, 2019. https://doi.org/10.1186/s12859-019-3080-2