Journal of Information Science and Engineering 38, 619-644 (2022)
DOI: 10.6688/JISE.202205_38(3).0008

A Multi-Dimensional Source Selection Based on Topic Modelling

Fatma Zohra Lebib[1,2], Hakima Mellah[1]
and Abdelkrim Meziane[1]
[1] *Information System and Multimedia System Division*
Research Center on Scientific and Technical Information
Algiers, 16028 Algeria
[2] *Department of Computer Science*
University of Sciences and Technology Houari Boumediene
Algiers, 16111 Algeria
E-mail: {zmatouk; hmellah; ameziane}@mail.cerist.dz

Access to information in multisource environments is facing many problems. One of them is the source selection problem. As more and more sources become available on the internet, how to select the relevant sources that meet the user needs is a big challenge. In this paper, we propose a multi-dimensional source selection approach based on topic modelling, which integrates both the social dimension and the intelligent dimension in order to optimize the source selection according to different user interests. Social tagging data is analyzed to discover relevant topics of user interests and latent relationships between users and sources based on topic modelling. By intelligently exploring a large search space of possible solutions, an (optimal) selection of sources is found using an intelligent method (a genetic algorithm). The proposed approach is evaluated on real data sources. The experimental results demonstrate that the proposed approach outperforms state-of-the-art source selection algorithms.

Keywords: multisource environment, social tagging, source selection, genetic algorithm, LDA

1. INTRODUCTION

In information retrieval, the paradigm *"one size fits all"* means that the same results are provided to the same queries regardless of the users who issued them: the system delivers information which strictly satisfying the criteria of the query. However, different users may have different interests and information needs even though using the same query. For example, a computer scientist may query "apple" to find information about a computer brand, while nutritionist may use the same query to find a description of fruit. When such a query is issued, the system returns a list of results that mix different topics for all users, even if they have different interests.

Received January 27, 2021; revised October 31, 2021; accepted December 16, 2021.
Communicated by Berlin Chen.

The query alone does not represent the actual need for information for a given user. It is important to understand the user's needs by exploiting other sources of information such as user behaviour with the system and user interaction with other users, which allow search results to be tailored to her/his areas of interest.

In a multisource environment, the relevant information is distributed within several locations or information sources. When searching in such an environment, it is crucial to select only relevant sources for a given user's query, filtering out irrelevant sources and searching only those that are likely to contain relevant documents [1]. This is referred in the literature as *source selection*, which is one of the main phases of a multisource information retrieval system [2].

Three major research problems are considered a multisource information retrieval system [2,3], which are: source representation [4], source selection [5,6] and result merging [7, 8]. Source representation gathers important information about the sources such as their contents and their sizes. Source selection selects a subset of information sources which are most relevant for a given query. Result merging combines documents retrieved from selected sources into a single ranked list before presenting the list to the end users. Source selection is a critical function of a multisource information retrieval system in which the broker attempts to route queries only to those sources which (potentially) contain relevant information, the source selection is then the problem we are addressing in this paper.

Existing source selection methods typically focus only on the query terms and the content of the source, while ignoring the user who submitted the query and his/her interactions with other users. Social networks have become an integral part of users' live for sharing and disseminating information. Users share their perspective on a topic or event and receive feedback, recommendations from peers, friends, *etc.*

In a multisource environment, the search for information is increasingly user-oriented and centered. In such systems, the main objective is the total satisfaction of the user. To obtain results close to the user, the search for information tends to model the user according to a profile (set of preferences) then to integrate it into the information access chain. In a multisource environment, in addition to taking into account the user profile which can be explicit or implicit (extract from his request), it would be even more interesting to propose an optimal selection of the sources to be interrogated according to this profile which would be involved in a social relationship.

In this paper, we propose a multi-dimensional source selection approach based on topic modelling that integrates both the social aspect and the intelligent aspect in order to select the best (optimal) selection of sources that correspond to the user, when the latter has a profile in relations with other users in social networks.

The rest of the paper is structured as follows. Section 2 presents the research problem and its objectives. Section 3 reports on prior work. Section 4 describes the proposed approach. Section 5 describes the conducted experiments. Section 6 reports and discusses the results obtained and Section 7 concludes the paper.

2. RESEARCH PROBLEM AND OBJECTIVES

The main problem that we are facing is the source selection when the required information is expressed in a multisource environment. When the search space, defined by the number of available information sources, is very large, it is important to explore this space intelligently to identify a small set of relevant sources that match the user interests.

Social environments can be exploited for various purposes, such as discovering user interests which are the key elements for improving adaptation [9]. User behaviour through the use of information sources can be useful to understand the sources more or less interesting for him. In a social tagging system, the user can annotate the sources he/she uses with a set of tags. Each source and each user can then be associated with a cloud of tags. The mega data generated from user tags can also contain many synonyms and polysemias, making it difficult to analyze them to understand the different user interests. One solution is to group these tags into "topics of interest" or "clusters", consisting of associated sets of tags sharing a semantic meaning in each cluster. There are some noticeable extensions of basic clustering methods, such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) [10]. These methods use adapted probabilistic models to cluster the content. The main assumption of these techniques is the existence of some implicit topics or purposes. They allocate all the content to these topics [11].

To solve the above problems, we propose a multi-dimensional source selection approach based on topic modelling that integrates both the social aspect and the intelligent aspect into the source selection process. The social dimension is the consideration of social relationships between users in social networks to adapt the source selection to the user interests, and the intelligent dimension is the use of artificial intelligence methods to optimize the source selection in order to restrict the sending of the query to an optimal selection of sources containing the relevant information. The topic modelling technique (LDA) [10] is used to discover latent relationships between users and sources from social tag data. The proposed approach is compared to state-of the-art personalized and non-personalized source selection approaches using differnt metrics to demonstrate its effectiveness on real datasets.

3. RELATED WORK

Generally, the goal of the search, in a multisource environment, is to send queries to as few sources as possible, therefore a source selection algorithm is applied. During the last two decades many approaches have been proposed for source selection [12–15], which can be grouped in three important categories: Big-document based approach, Small-document based approach and Classification-based approach. Big-document approaches view sources as large documents represented as a bag of words, where the document ranking algorithms are adapted to rank sources, such as CORI [16], GlOSS [13], CVV [17] and recently Taily [18].

Small-document approaches analyze individual sample documents in the source representation to rank sources according to the ranking of their sampled documents for a query, like ReDDE [15], CRCS [14], SUSHI [6, 19]. ReDDE selection algorithm ranks sources based on their expected number of relevant documents. A recent survey by Markov and Crestani [20] provides a detailed analysis of small-document approaches. Classification based approaches treat source selection as a classification problem [12, 21–25]. A classification model can be learned from a set of training queries and is used to predict the relevance of a source for test queries.

In addition, some source selection approaches [25–27] have used learning to rank methods to classify sources, and others [28,29] have investigated the efficiency of source selection using appropriate strategies such as load balancing methods.

Most of the previous research has focused on assessing the relevance of a source by analyzing its static information [22]. Recent works consider other important information such as: the past queries results [12], the results diversification [30], the importance and trustworthiness of sources and results [31, 32], the relevance and novelty [33], the source quality [34–38], the query context [39] and the semantic meaning representations and semantic distance between the query and the source [40].

The source selection problem can be viewed as a combinatorial optimization problem [41] , which consists of finding the "best selection" (solution) by exploring a large research space of possible solutions using a Genetic Algorithm (GA).

Few works take into account the information characterizing the user to tackle the problem of source selection in multisource environments. Lu and Callan [42] proposed an approach for modelling user interests to improve the efficiency of full-text federated search in peer-to-peer networks. Their approach models a user's persistent, long-term interests based on past queries, and uses the model to improve search efficiency for future queries that represent interests similar to past queries. Kechid and Drias [43] proposed an approach for personalizing information retrieval in a distributed environment based on multi-agents technology. Their approach consists in the integration of both user profile and source profile in the source selection process and the results merging steps. The user profile is built using personal data, search history, and preferences. Recently, in [44], The authors proposed a user-centered approach that addresses source selection as a multicriteria problem. Their approach uses a decision support methodology to allow users to formalize their preferences by specifying the relative importance of the different criteria. User preferences are then used in an optimization framework to find the most appropriate sources for the user.

Social data can be exploited to improve personalized search [45–48] and to personalize recommendation [49], by modeling the user's profile from social tagging. User-generated tags are effective in representing user interests because these tags reflect the judgments of human beings while being more concise and closer to human comprehension [50]. The latent structure between users, tags and resources [51] allows to discover the social interests of users [52] and therefore to ensure user satisfaction in information retrieval.

In [53], the social user profile is used to personalize and improve the distributed information retrieval. A folksonomy structure between the three entities (users, documents, tags) is exploited for enhancing query expansion and improving both the source selection and the result merging. The user profile is created by the set of tags. The authors took into account the user's short-term interests based on his/her current query. In their work, it is not considered the multiple topics of interest of the user and their impact on the problem of source selection. The user may have a mixture of topics of interest, to discover and to extract from a large volume of data characterizing the user.

Topic modelling [54,55] provides methods for latent knowledge discovery, finding relationships among data, understanding, and summarizing huge corpus of data [56]. Latent Dirichlet Allocation (LDA) [10] is one of the most popular methods in topic modelling for discovering latent semantic topics in large collections of text. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Topic models are applied in various fields, including information retrieval [57–61] and recommender systems [62–64].

In [57], authors studied the use of LDA to improve ad-hoc retrieval. They proposed an LDA-based document model within the language modelling framework, and evaluated it on several TREC[1] collections. They showed that improvements over retrieval using cluster-based models can be obtained with reasonable efficiency. Baillie *et al.* [58] proposed a collection selection approach that models the collection in a low dimensional topic space. They presented an extended version of latent Dirichlet allocation that uses a hierarchical hyperprior to enable the different topical distributions found in each collection to be modelled. Under the model, resources are ranked based on the topical relationship between query and collection.

In [59], a model called the Personalized Topic Model (PTM) was presented for personalized search from query logs using sets of latent topics derived directly from the log files. User profiles are constructed based on the representation of clicked documents over a topic space. Their experiments showed that by subtly introducing user profiles as part of the ranking algorithm, rather than by reranking an existing list, it can provide personalized ranked lists of documents which improve significantly over a non-personalized baseline. A dynamic topic model that personalizes dynamically the information retrieval is proposed in [60]. The proposed model predicts dynamic user interest based on query log and addresses the challenging problem of predicting results for new users.

Carman *et al.* [61] investigated the utility of topic models for the task of customizing search results based on information present in a large query log. They described two different topic models capable of factoring the query log into a set of parameter matrices, and define document ranking functions based on the learned parameters. They extend Latent Dirichlet Allocation (LDA) topic model by adding user variables in order to create more complex models for query log analysis.

For recommendation system, Zhao *et al.* proposed a personalized hashtag recommendation approach based LDA model that can discover latent topics in microblogs [63]. Giri *et al.* proposed an unsupervised topic model to understand

[1] https://trec.nist.gov/

the interests of the cellular users based upon their browsing profile. The proposed model allows extracting users' hidden interests, which can be very effective for mobile advertisements and recommendation systems [62]. Jin *et al.* focused on the issue of tag recommendation. They proposed hybrids approach based on a combination of Language Model (LM) and LDA for tag recommendation in terms of topic knowledge [64]. In [65], authors proposed three generative probabilistic models to represent users' interests in a latent space over resources and tags used to describing them. They showed that latent user interests combined with social clues from the immediate neighborhood of users can significantly improve social link prediction in the online music social media site Last.fm. Liao *et al.* proposed model that allows to discover latent event topics from event text descriptions with a latent Poisson topic model. Associations between latent topics and participant influence are exploited to improve event recommendation [66].

To summarize, few works take into account user information in the source selection in multisource environments, this is due to the lack of central control in such a complex environment composed of a huge number of sources and user profiles. In this work, we study the problem of source selection in a multisource environment by exploiting emerging technologies such as evolutionary algorithms and machine learning to improve research on these environments. Evolutionary and meta-heuristic algorithms have been widely applied to solve several optimization and search problems. They can obtain near-optimal solutions within a short period of time; thus, it has attracted the attention of academia in recent years. Emerging social networks have involved large amounts of data generated with complex structures. Big data generated from these users and sources is typically network-based, which may reflect the relationship between users and sources. Analyzing this big data requires machine learning methods in order to learn the semantic relationships between users and sources and therefore improve the accurancy of source selection approaches.

4. THE PROPOSED APPROACH

A new approach to source selection that combines two dimensions, namely social and intelligent, is proposed. This approach aims to find the optimal selection of sources corresponding to the topics of interest of a given user. LDA topic modelling is used to represent user interests on a reduced-dimensionality latent topic space. LDA, which is traditionally used to model textual content, is used to find patterns of user behaviour in social tagging and group them accordingly. This means that the topics space is extracted directly from all the tags used by users to annotate the available information sources. The discovered topics allow to understand the interests of the user and thus to customize the source selection. An artificial intelligence method is used to find the optimal solution to the source selection problem by considering the topics of interest of each user. First, we try to tackle the problem with a genetic algorithm for its simplicity of implementation, but other artificial intelligence algorithms can also be used.

4.1 Problem Definition

Before introducing the detailed description of the proposed approach, a formal definition of the source selection problem is given as follows:

We consider, $S = \{s_1, s_2, ..., s_n\}$, a set of n information sources, $|S| = n$.

$T(u_i) = \{t_1, t_2, ..., t_m\}$, a set of m tags used by user u_i to annotate sources.

$q(u_i) = \{k_1, k_2, ..., k_p\}$, a set of p independent query's terms of user u_i.

Question: find a subset S' of S for a specific user, such as the similarity between the elements of S' and the pair (query, user's topics of interest) is maximal, where $|S'| = k$. k is the number of selected sources, where $k < |S|$.

E is the search space of possible solutions consisting of k-combinations, $|E| = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

When n is very large, the number of possible combinations is enormous and no complete method is able to yield a solution of good quality. One approach to cope with this issue is the use of artificial intelligence techniques such as a genetic algorithm.

A solution to this problem is a selection of k-sources called "*a selection*" denoted **sel**.

In this paper, we assume that an information source is considered as a large document represented by a set of terms extracted from its sampled documents [13,16,17].

4.2 Description of the Approach

The proposed approach includes two major steps, as shown in Fig. 1. The first step is "Users' topics of interest discovery" which consists in discovering the topics of interest from a set of tags used by users to annotate the available sources, for this, the LDA model is used. The output of the LDA algorithm consists in the probabilities distribution over all the topics of interest, for each user and each source. The output of LDA is the input of the second step, which is "Source selection process". In this step, the best selection of sources that maximizes the fitness function is generated using a genetic algorithm. The solution is a set of k sources that satisfy a given user. These two steps are described in detail below. Note that LDA topic modelling is an off-line step that is performed independently of GA.

4.2.1 User's topics of interest discovery

With the use of an information source, users can annotate sources (or documents of sources) using a set of tags. These tags are seen as an important data source to capture the interests of users. LDA-based topic modelling is used to infer a user's overall topical interest by combining all user tags. Each tag is associated with a particular document or a source. This association can directly find the topic (s) to which a source is linked, which help to deduce the topics from the sources.

By using LDA, a user is considered as "a document" and the set of tags used by this user as the "words" contained in this document. As such, each topic

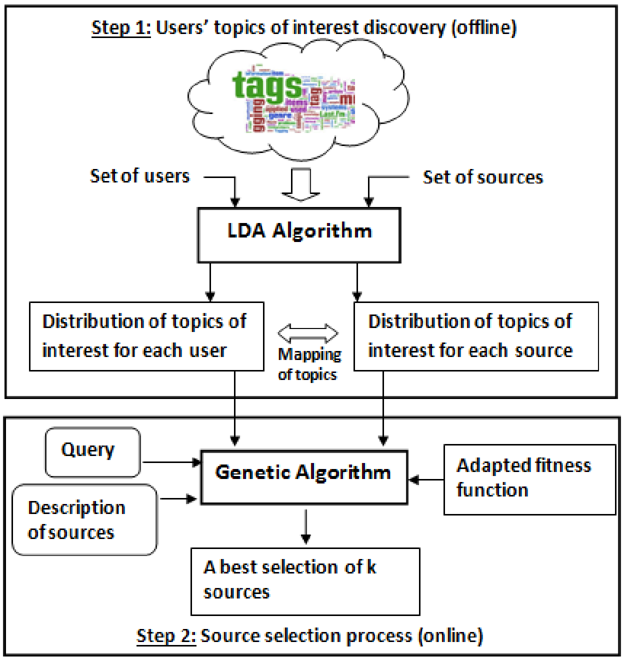

Step 1: Users' topics of interest discovery (offline)
tags
Set of users
Set of sources
LDA Algorithm
Distribution of topics of interest for each user
Mapping of topics
Distribution of topics of interest for each source
Query
Genetic Algorithm
Adapted fitness function
Description of sources
A best selection of k sources
Step 2: Source selection process (online)

Fig. 1. Description of the proposed approach.

discovered can be interpreted as "a topic of interest" of the user. The LDA modelling process consists to find a mix of topics of interest for each user from the tags posted by all users for the sources, this ensures greater accuracy.

Suppose there are Z topics of interest that we would like to explore, and U users. We de note by $T(u_i)$ the set of tags used by the user u_i to annotate the available sources.

$T(u_i) = \{t_1, t_2, ..., t_m\}$, $u_i \in U$ and $|T(u_i)| = m$

The tags of all users produce a vocabulary which is then used to generate the latent topics of interest. The output of the LDA model provides, for each user u_i, a vector of probabilities distribution over Z topics of interest, noted as follows:

$D(u_i) = \{w_1, w_2, ..., w_z\}$, $u_i \in U$ and $|D(u_i)| = Z$, it represents the distribution of the topics of interest of a user, where w_i is the user's interest rate relative to i^{th} topic.

The same vocabulary (the tags of all users) is used to learn the topics for a source from the set of tags related to that source. Each source s_j represented by a collection of tags is mapped to generated topics that are represented by tag distributions, to infer the topics related to s_j. An LDA technique called *inference* is used to generate the probabilities distribution over Z topics of interest for the source s_j, noted as follows:

$D(s_j) = \{v_1, v_2, ..., v_z\}$, $s_j \in S$ and $|D(s_j)| = Z$, it describes the importance of a source to each topic, where v_i is the belonging rate of the source to i^{th} topic.

Topic modelling allows to model topic-user and topic-source relationships.

Thus, a hidden relation (a connection) between the user u_i and the source s_j can be deduced, allowing to identify the sources close to the user's interests according to their similarity. To compute the interest of user u_i for the source s_j (denoted, Interest (u_i,s_j)), the cosine formula [67, 68] is used. It calculates the similarity between the two vectors $D(u_i)$ and $D(s_j)$ given in Eq. (1).

$$Interest(u_i, s_j) = Similarity(u_i, s_j) = \frac{\sum_{h=1,Z}(w_{ih} * v_{jh})}{\sqrt{\sum_{h=1,Z}(w_{ih})^2 * \sum_{h=1,Z}(v_{jh})^2}} \quad (1)$$

4.2.2 Source selection process

In this step, a genetic algorithm is applied to generate the best solution based on the source description, the user's query and the user's topics of interest. We present below the genetic representation of the potential solutions (Chromosomes), the fitness function and the genetic operators, followed by the proposed genetic algorithm.

- **Chromosomes Representation.** A solution to the defined source selection problem is a combination of k sources. Each source is encoded, in a simple way, by an integer between 1 and n (n: number of available sources). Therefore, a chromosome, which is an input to the genetic algorithm, is represented by a vector of length k.
  The initial population of the genetic algorithm is generated randomly from the search space E consisting of possible k-combinations.
  Example:
  Let consider 8 information sources ($n = 8$) and the number of sources to be selected is equal to 6 ($k = 6$).
  $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $|S'| = 6$
  $|E| = \binom{8}{6} = \frac{8!}{6!(8-6)!} = 28$.
  Examples of possible solutions are presented by the chromosomes below, avoiding the same gene value into the chromosome.
  Chromosome1: 1 2 3 4 5 6
  Chromosome2: 5 2 3 4 6 7
  Chromosome3: 1 2 3 5 7 8
- **Fitness function.** A solution to the problem is denoted by (sel). To assess the relevance of sel, the relevance average of the k sources that appear in sel is used, given by Eq. (2).

  $$Relevance(sel, u_i, q) = \frac{\sum_{j=1}^{k} Relevance(s_j, u_i, q)}{k} \quad (2)$$

  Where,
  Relevance(s_j, u_i, q) : the relevance value of the source s_j for the query q and the user u_i
  k: the number of selected sources, $k = |sel|$
  The relevance evaluation of a source for a query and a user considers the triplet (source, user, query). It is given by the following equation.

  $$Relevance(s_j, u_i, q) = (1 - \lambda)similarity(s_j, q) + \lambda Interest(u_i, s_j) \quad (3)$$

Where,
$Similarity(s_j, q)$ is the similarity between a source and query.

The query and the source are represented by vectors of terms weights in an M-dimensional space corresponding to the M terms present in the search space. Thus the similarity between a source s_j and a query q is calculated by the cosine formula between the two vectors of terms weights of the source and the query respectively in Eq. (4).

$$Similarity(s_j, q) = \frac{\sum_{i=1}^{M}(wt_{ji} * wt_{qi})}{\sqrt{\sum_{i=1}^{M}(wt_{ji})^2 * \sum_{i=1}^{M}(wt_{qi})^2}} \tag{4}$$

Where, wt_{ji} and wt_{qi} are the weights[2] of the term i in the source s_j and query q respectively.

$Interest(u_i, s_j)$ represents the interest of the user u_i for the source s_j calculated by Eq. (1).
The parameter λ, in the range zero to one, controls the effect of the user's topics of interest on the overall evaluation

- **Genetic operators.** The genetic algorithm uses its genetic operators with configurable probabilities to generate offspring based on the current population. Once a new generation has been created, the genetic process is repeated iteratively until a solution of problem is found or the maximum number of generations is reached.

Selection. It simulates the "survival-of-the-fittest". Selection replicates chromosomes (solutions to the problem) with high fitness values and removes chromosomes with low fitness values. We used natural selection which takes the best chromosomes for the next generation. The best chromosomes are identified by evaluating their fitness value.

Crossover. It combines two chromosomes together to form new offspring. Crossover occurs only with a crossover probability P_c. The chromosomes are not subjected to crossover remain unchanged. Crossover allows exploiting the current solution to find better chromosomes. The single-point crossover is used, exchanges the values of sub-vector between two chromosomes, which are candidates for this process. Algorithm 1 describes the crossover operator.

Mutation. It is the process of randomly altering the genes in a particular chromosome. The mutation involves changing the gene values of a solution with certain probability P_m. The mutation allows the exploration of the whole search space. We used a single-point mutation, in which only one gene is changed. The mutation operator is described by Algorithm 2.

[2] The weight of a term is calculated using the $tf - idf$ approach.

Algorithm 1: Crossover Algorithm

```
Input: two candidate chromosomes X and Y.
Output: two new chromosomes $\overline{X}$ and $\overline{Y}$.
Let $X = (x_1, x_2, ..., x_k)$ and $Y = (y_1, y_2, ...y_k)$.
begin
    Choose a random gene along the length, at the position $p$, and swap all
    the genes after that point. Two new chromosomes $\overline{X}$ and $\overline{Y}$ are created
    according to the following rule:
        $\overline{x} = \begin{cases} x_i & if\ i < r \\ y_i & otherwise \end{cases}$
        $\overline{y} = \begin{cases} y_i & if\ i < r \\ x_i & otherwise \end{cases}$
    Remove, before the cutting point ($p$), the sources which are already
    placed after the cutting point.
    Identify sources that do not appear in each of the two chromosomes.
    Randomly fill the holes in each chromosome.
end
```

Algorithm 2: Mutation Algorithm

```
Input: the current population ($C$).
Output: the population at the next generation ($N$).
Let $p_m$: mutation probability.
begin
    for each chromosome in $C$ do
        Generate a random number $r$ on the interval $[0, 1]$.
        if ($r < p_m$) then
            Apply the mutation operator to this chromosome:
            begin
                Choose a random gene along the length, at the position $p$,
                alter the value of the gene $p$ with a new value that does not already
                exist in the chromosome (the new value to be placed is generated randomly
                between 1 and $n$).
            end
            Insert the mutated chromosome into $N$.
        endif
        else insert the chromosome into $N$ without change.
    end for
end
```

- **Genetic algorithm.** The source selection is adapted to each user, by considering her/his topics of interest. The proposed genetic algorithm aims to find the set of sources that best matches the user interests. The fitness function that evaluates the performance of each solution linearly combines the similarity between the query and the source, as well as the similarity between the user and the source. Thus, for the same query, the best selection may be different for two users with different interests. The proposed algorithm is called Personalized Genetic Algorithm for Sources Selection based LDA (Algorithm 3).

Algorithm 3: PGASS-based-LDA: Personalized Genetic Algorithm for Sources Selection based LDA

```
Input: $n$ sources, topics of interest of a user $u_i$ and the query $q$.
Output: $k$-best information sources for the user $u_i$.
Let,
$p_m$ : mutation probability
$p_c$ : crossover probability
$Pop_{Size}$ : population size
$Gen_{Max}$ : maximum number of generation
begin
1: Generate randomly an initial population of $Pop_{Size}$ size from the possible solutions.
2: Evaluate each solution in the initial population using the fitness function given by Eq. (2).
3: Genetic Evolution:
while termination criterion not reached ($Gen_{Max}$ is not reached) do
  a. Select the appropriate chromosomes for reproduction.
  b. Apply the crossover operator to the pair of parents according to $p_c$ to produce new chromosomes (offspring).
  c. Add the new offspring to the population.
  d. Apply the mutation operator for each Chromosome in the population according to $p_m$.
  e. Add the modified chromosomes to the population.
4: Evaluate the current population using fitness function given in Eq. (2).
5: Select the best performing chromosomes for the next generation (use newly generated population for a further run of the algorithm).
6: end while
end
```

4.3 Issues of the Proposed Source Selection Approach

The proposed approach uses social tagging data to infer the topics of interest of users using LDA. However, when this data is not available, for a new user or for sources those are not tagged by users, which are the well-known issue of data sparsity and cold start problem. In this case, we suggest exploiting user-user relationships using previous queries to identify the most similar users. The proposed approach can then use the distribution of topics of interest of the top similar users for the source selection.

For a new user, to overcome this limitation, the proposed approach uses the vocabulary containing the words composing all the users' queries and proceeds as follows:

- Calculate the probabilities distribution of existing users over the query elements using LDA.
- For each new user, calculate his/her distribution over the query elements by considering all the terms of queries (the same vocabulary size).
- Calculate the similarity between a new user's distribution and each distribution of all existing users using cosine formula.
- Select the user more similar to the new user.
- Select sources for the new user using his/her query and the distribution of topics of interest of the selected user (his/her similar).

We can also explore other social tagging datasets, such as del.icio.us dataset for user profile enrichment. Note that the demonstration of these solutions is not discussed in this paper.

5. EXPERIMENTS

This section describes the experiments conducted to evaluate the proposed approach. The proposed personalized approach is compared with personalized and non-personalized approaches chosen from the state-of-the-art source selection algorithms. First, details about baselines are given, then the datasets and the metrics used in the evaluation of the approaches are presented with the description of the execution of LDA modelling to generate the users' topics of interest needed for the PGASS-based-LDA algorithm and finally the results and the discussions are presented.

5.1 Baselines

The performance of the proposed approach is compared with four state-of-the art models, one personalized and others no personalized.

- The most popular CORI algorithm [16] which views each collection or source as one large document, is considered as one of the most stable and effective source selection algorithms [69]. CORI is based on Bayesian inference networks. In CORI the similarities between a user query and a set of document collections is computed, in order to rank the collections.
- A vocabulary-based source selection algorithm, Taily [18] that models a query's score distribution in each shard or source as a Gamma distribution and selects shards with highly scored documents in the tail of the distribution. Taily estimates the parameters of score distributions based on the mean and variance of the score function's features in the collections and shards. Taily algorithm uses two parameters, n_c and v, where n_c is roughly the depth of the final ranked list desired, and v is the number of documents in the top n_c that a shard must be estimated as contributing in order to be selected. The parameters values ($n = 400$ and $v = 50$) recommended by Aly *et al.* are used in the experiments.
- The GASS algorithm [41] uses a genetic algorithm for the selection of sources, but does not consider the social aspect.

- A personalized source selection algorithm [53] (denoted SaoudAlgo), this approach integrates social profile information into the source selection process. Social tagging data is used to create a social profile of each user. sources are ranked according to a score that combines the similarity between the source and the query according to the set of terms of the source documents, and the similarity between the source and the user according to the set of tags of the source documents.

The genetic algorithms are implemented in a java environment using the java genetic algorithm library JGAP[3]. The same parameters values presented in Table 1 are used for the two approaches based genetic algorithm, namely GASS and PGASS-based-LDA.

Table 1. The parameters values of genetic algorithms.

Parameter	value
Crossover rate	60%
Mutation rate	10%
Generation number	500
Population size	50

5.2 Datasets

5.2.1 Information source datasets

Real information sources are used in the evaluation of source selection approaches which are online databases of scientific research articles from different areas such as computer science, economics, finance, *etc.* The access to these sources is through SNDL[4] platform. Table 2 describes the information sources used in the expirments. We consider an information source as a large document each source

Table 2. SNDL datasets used in the experiments (8 information sources).

Source number	Information source	Source domain
1	ACM Digital Library	Computer science
2	Edward Elgar Products	Economics, finance, business and management, law, *etc.*
3	IEEE	Computer science, Electronics, Telecommunications
4	IOP science Extra	Physics, Materials Science, Applied Mathematics
5	JSTOR	Multidisciplinary
6	Royal Society of Chemistry	Chemistry, Materials Science, Environment, Biology
7	ScienceDirect of Elsevier	Multidisciplinary
8	SpringerLink	Multidisciplinary

[3] JGAP is a Genetic Algorithms and Genetic Programming framework written in Java (http://jgap.sourceforge.net/).

[4] https://www.sndl.cerist.dz

is represented by a set of terms extracted from its sampled documents. To create the sources description, we used a user account in the SNDL platform, which allow us to search and download documents using probe queries. For that, query-based sampling thechnique [4] was used. 15 one-word queries are send to each source. The queries terms are chosen from the most popular tags. We downloaded the first 4 documents for each query, to obtain approximately 60 documents per source. The above process may not produce optimal representatives as noted by Thomas and Hawking (2007), but has become standard practice when evaluating source selection algorithms [14, 15]. The information sources are indexed with Indri [70] which is an open-source indexing and information searching system from Lemur Project [71]. A common index that group terms from all sources is built. Then the index file is cleaned te remove non-significant terms, this is done manually. The vectors of sources and test queries are constructed using the $tf * idf$ approach.

5.2.2 Social datasets

The 2017-01-01 version of the BibSonomy datasets is used. The dataset contains public bookmarks and publication posts of Bibsonomy. Only the Bibtex file containing information about BibTeX data (scientific publications) is used. The dataset has been created using the mysqldump command of a MySQL database, which makes easy their manipulation. The dataset are filtered in order to extract only the data concerning the sources used in the experiments, sources of which we have built an accurate description for testing purposes. The resulting reduced dataset is described in more detail in Table 3, where the items are the documents from the sources (Table 1) and which are tagged by the users.

Table 3. Social datasets features.

BibSonomy dataset			
Users	Distinct Items (URL)	Individual Tags	Distinct Tags
2807	95014	254732	38291

Table 4 summarizes for each source the number of associated tags.

Table 4. Number of tags associated with sources.

BibSonomy dataset		
Source Number	Source Name	Number of Tags
1	ACM	70387
2	Elgar	421
3	IEEE	33628
4	IOP	236
5	JSTOR	6431
6	RSC	958
7	Sciencedirect	135338
8	Springerlink	7333

5.3 Evaluation Metrics

Metrics for evaluating source selection methods are usually based recall and precision. Due to unavailable of the relevance judgments concerning the total number of relevant documents available in the information sources used in the experiments, the evaluation based precision is used to evaluate the proposed approach. The source-level precision is used to evaluate the performance of the source selection approaches, given by the following equation.

$$Precision = \frac{number\ of\ relevant\ sources}{number\ of\ selected\ sources\ (k)} \tag{5}$$

Two other metrics are also used to evaluate the proposed approach, the Mean Average Precision (MAP) which is the average precision over multiple queries/rankings, and the Mean Reciprocal Rank (MRR) which shows how the best relavant sources are ranked in a high position.

The relevant of a source s_j depends on pair (u_i, q) (user u_i submits the query q), that is the personal relevance judgments required to evaluate any personalized source selection approach. To construct these relevance judgments, we used social data and we posited the following hypothesis: any source s_j labeled by u_i with at least one term of q is considered to be relevant for the pair (u_i, q). These relevance judgments require considerable effort to generate the test queries and labelling relevant sources. The purpose of this evaluation is to verify, for each query and each user, whether a source labeled as relevant (*i.e.* the source that the user has annotated using the terms of the query) appears in first of his/her final result as a good solution. The five algorithms CORI, Taily, GASS, SaoudAlgo and PGASS-based-LDA are compared using test queries consisting of 2 to 6 terms each and a number of users selected from the social dataset. The average precision of these four algorithms is calculated for 15 users and 12 test queries, for each user and each query the precision given by Eq. (5) is calculated then the average over 12 test queries is calculated before calculating the precision final average over 15 users. Note that the test queries and users are are carefully chosen in order to be able to show the improvement of the proposed approach. Users are selected from the social dataset used in the experiments, and queries are generated taking into account the content of the sources and user tags.

5.4 Performing LDA Topic Modelling

Java implementation of LDA (JGibbLDA[5]) [72] is used to generate hidden topics of interest of users. This implementation relies on Gibbs sampling to learn the distributions, which requires the following parameters. Z: number of topics, *beta* and *alpha*, the Dirichlet priors, and N, number of iterations.

We set the default values for hyper-parameters, $\alpha = \frac{50.0}{Z}$ and $\beta = 0.1$, where Z is the number of topics considered ($Z = 100$). For all experiments, the LDA operations are run through 1000 iterations of Gibbs Sampling. We obtain the distribution of topics of interest to each user.

[5] JGibbLDA, http://Jgibblda.sourceforge.net/

Using the previously estimated LDA models, the probabilities that a source belongs to each of the topics of interest are inferred using a technique called *inference*. The source then can be represented with a distribution vector over the set of generated topics of interest.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In the experiments we varied the number of selected sources (k = 2, 4, 6, 7) among the eight (8) sources. The proposed approach is evaluated on 15 users and 12 test queries, which present 12*15 = 180 cases to be evaluated. The same users and queries test are used to evaluate the baseline models. In this section, we analyze and discuss the following points: (1) impact of the parameter λ on the performance of PGASS-based-LDA algorithm; (2) comparison of performance of the proposed approach with four baselines; and (3) time complexity of the proposed approach compared to the baseline models.

6.1 Impact of the Parameter λ

The parameter λ of PGASS-based-LDA algorithm is used to assess the fitness of each solution in Eq. (3). It shows the impact of incorporating topics of user interest when assessing the relevance of a source. By varying the values of λ in the range $[0, 1]$, we can deduct if the relevance of the source is more or less related to the user interests or to the query. If only sources close to the user's interests are taken into account in the evaluation ($\lambda = 1$), then the result can be a small or empty set while good results can be found in others sources that are seldom or rarely used by a user, their importance is not perceived by the user even if they are relevant to his/her request. We tested the impact of this parameter with the values $\lambda = 0.1$, $\lambda = 0.4$ and $\lambda = 0.55$. Table 5 shows the performance of the proposed approach in terms of precision for $k = 2, 4, 6$ and 7. When $\lambda = 0.1$, the proposed approach provides high performance over the datasets (see Fig. 2). We put $\lambda = 0.1$ to compare the source selection approaches.

Table 5. Average precision of PGASS-based-LDA on the SNDL datasets (by varying λ).

λ	Number of selected sources			
	k=2	k=4	k=6	k=7
λ=0.1	0.5778	0.4417	0.3250	0.2802
λ=0.4	0.5722	0.4111	0.3185	0.2794
λ=0.55	0.5417	0.4028	0.3158	0.2762

6.2 Comparison of Different Source Selection Approaches

Table 6 shows P@{2,4,6,7}, MAP and MRR of the five source selection algorithms. The results over the SNDL datasets indicated that the proposed algorithm (PGASS-based-LDA) offers the best performance compared to the other four algorithms (Fig. 3). This shows that taking user information into account in the source

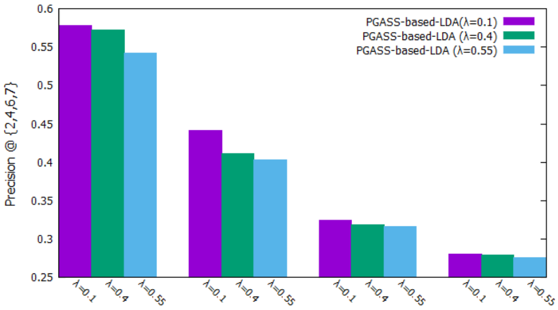


Fig. 2. Impact of the parameter value λ on the PGASS-based-LDA algorithm performance.

selection process improves the accuracy of results compared to non-personalized solutions (CORI, GASS and Taily). And compared to the personaized approach (SaoudAlgo), the proposed approach is more efficient than the SaoudAlgo approach, this is due to the modelling of the topics of user interest from the social tag data instead of using directly the raw tag data . The results also indicated that PGASS-based-LDA was more efficient in terms of MRR which explains that the most interesting sources for the user move up in the rankings, they change position compared to the initial positions obtained by using the four other algorithms (CORI, Taily, GASS and SaoudAlgo).

Table 6. Result comparison of different algorithms over the SNDL datasets.

Source selection algorithm	P@2	P@4	P@6	P@7	MAP	MRR
CORI	0.3305	0.3722	0.2898	0.2635	0.3140	0.6994
Taily (n=400, v=50)	0.3861	0.3555	0.2917	0.2643	0.3302	0.7662
GASS	0.4639	0.3805	0.3148	0.2762	0.3588	0.8672
SaoudAlgo ($\alpha = 0.1$)	0.45	0.3839	0.3148	0.2769	0.3577	0.7784
PGASS-based-LDA (λ=0.1)	0.5778	0.4417	0.3250	0.2802	0.4061	0.9123

6.3 Time Complexity of Proposed Approach

The genetic algorithm runs in iterations (or generations). Initially, a set of solutions are generated randomly (called a population). Crossover and mutation operations are done over the solutions in each of iterations. The best k solutions are kept in population for the next iteration. After the last iteration, the best solution is found. We note here that the time cost of an iteration depends on the inner operations (*e.g.* crossovers, mutation, generate random solutions, *etc.*) which are usually simple to implement, and also problem-dependent. In general, they depend on the size of a solution. The execution time of a genetic algorithm also depends on the number of iterations [73]. Typically, we want to stop when we converge to a solution that is hardly improved. How to find the number of iterations that

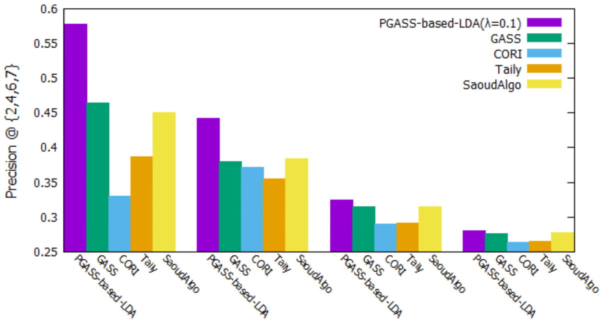


Fig. 3. Comparison of source selection algorithms on SNDL datasets.

guarantee this some probabilistic analyses to find the average convergence time [74]. Note that the number of interactions to reach the optimal solution can be fixed in the experiments.

During the experiments, we varied the iterations number of the proposed algorithm and we checked the obtained results for 12 test queries. The proposed algorithm converges toward the optimal solution when the algorithm reaches the number of iterations ($Num.Iter$) equal to 500, 100 and 50, in each of these cases, the time cost of the algorithm is calculated. The execution time is also calculated for other algorithms used in performance evaluation. Table 7 shows the average execution time of each algorithm taken to respond to a user query. The approaches based on the genetic algorithm (PGASS-based-LDA and GASS) are more complex than CORI, Taily and SaoudAlgo in terms of time but offer better performance in terms of the quality of the generated solutions (as shown in Section 6.2). A number of iterations equal to 50 provides a better performance in time complexity of the proposed algorithm.

Table 7. Time complexity of the five algorithms.

Algorithms	Execution time (seconds)		
	$Num.Iter = 500$	$Num.Iter = 100$	$Num.Iter = 50$
PGASS-based-LDA	102.43	20.86	10.72
GASS	107,16	21.36	10.92
CORI	0.042		
Taily	5.83		
SaoudAlgo	4.79		

7. CONCLUSION AND FUTURE WORK

In this paper, we discussed the application of intelligent methods and social data to improve the information retrieval efficiency in a multisource environment. We focus on how to meet user needs and tailor source selection accordingly. A multi-dimensional source selection approach based on LDA topic modelling is proposed to solve the source selection problem, taking into account both the user's topics of interest and a large number of available sources.

On the one hand, the users' topics of interest are deduced from the tags assigned to the information sources, which also makes it possible to determine the topics of a source. The generated topics provide therefore an effective way to bridge the gap between users and sources. And on the other hand, we have proposed a new formulation of the source selection problem, defined as a combinatorial optimization problem where a solution consists of k sources selected among the available sources.

The results of the experiments have shown that the proposed approach outperforms the personalized and non-personalized source selection approaches of the state of the art. Note that the performance evaluation is time consuming and requires personal relevance judgements which are subjective and depend on user and her/his social profile. In this paper we exploit social tagging to build the personal relevance judgements. Other sources of information can also be exploited for this purpose, such as user feedback, user-source ratings, interaction logs, which are content types introduced by different social platforms.

Experiments are needed for setting the parameters values of the genetic algorithm (population size, generation number) to optimize the proposed algorithm and make it more efficient. These settings depend on many other parameters such as the problem size and the search space. We plan to use large publicly test collections to show that the solution is scalable and efficient in a multisource environment.

REFERENCES

1. I. Markov, M. J. Carman, and F. Crestani, "Towards risk-aware resource selection," in *Proceedings of Asia Information Retrieval Symposium*, 2014, pp. 148-159.
2. F. Crestani and I. Markov, "Distributed information retrieval and applications," in *Proceedings of European Conference on Information Retrieval*, 2013, pp. 865-868.
3. M. Shokouhi and L. Si, "Federated search," *Journal of Foundations and Trends in Information Retrieval*, Vol. 5, 2011, pp. 1-102.
4. J. Callan and M. Connell, "Query-based sampling of text databases," *ACM Transactions on Information Systems*, Vol. 19, 2001, pp. 97-130.
5. G. Paltoglou, M. Salampasis, and M. Satratzemi, "Integral based source selection for uncooperative distributed information retrieval environments," in *Proceedings of ACM Workshop on LSDS for IR*, 2008, pp. 67-74.
6. P. Thomas and M. Shokouhi, "Sushi: scoring scaled samples for server selection," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 419-426.
7. M. Shokouhi and J. Zobel, "Robust result merging using sample-based score estimates," *ACM Transactions of Information Systems*, Vol. 27, 2009, pp. 1-29.

8. I. Markov, A. Arampatzis, and F. Crestani, "On cori results merging," in *Proceedings of the 35th European Conference on Information Retrieval*, Vol. 7814, 2013, pp. 736-739.
9. M. Mezghani, A. P. C. A. Zayani, I. Amous, and F. Sèdes, "Producing relevant interests from social networks by mining users' tagging behavior: A first step towards adapting social information," *Data and Knowledge Engineering*, Vol. 108, 2017, pp. 15-29.
10. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
11. J. Kim and J.-H. Lee, "A novel recommendation approach based on chronological cohesive units in content consuming logs," *Information Sciences*, Vol. 470, 2019, pp. 141-155.
12. S. Cetintas, L. Si, and H. Yuan, "Learning from past queries for resource selection," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 1867-1870.
13. L. Gravano, H. Garcia-Molina, and A. Tomasic, "Gloss: text-source discovery over the internet," *ACM Transactionsons on Information Systems*, Vol. 24, 1999, pp. 229-264.
14. M. Shokouhi, "Central-rank-based collection selection in uncooperative distributed information retrieval," in *Proceedings of the 29th European Conference on Information Retrieval*, 2007, pp. 160-172.
15. L. Si and J. Callan, "Relevant document distribution estimation method for resource selection," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 298-305.
16. J. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 21-29.
17. B. Yuwono and D. L. Lee, "Server ranking for distributed text retrieval systems on the internet," in *Proceedings of the 5th International Conference on Database Systems for Advanced Applications*, 1997, pp. 41-49.
18. R. Aly, D. Hiemstra, and T. Demeester, "Taily: Shard selection using the tail of score distributions," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 673-682.
19. A. Kulkarni, A. S. Tigelaar, D. Hiemstra, and J. Callan, "Shard ranking and cutoff estimation for topically partitioned collections," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 555-564.
20. I. Markov and F. Crestani, "Theoretical, qualitative, and quantitative analyses of small-document approaches to resource selection," *ACM Transactions on Information Systems*, Vol. 32, 2014, pp. 1-37.
21. J. Arguello, J. Callan, and F. Diaz, "Classification-based resource selection," in *Proceedings of the 18th International ACM Conference on Information and Knowledge Management*, 2009, pp. 1277-1286.

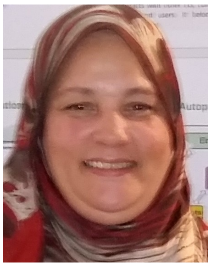
22. D. Hong, L. Si, P. Bracke, M. Witt, and T. Juchcinski, "A joint probabilistic classification model for resource selection," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010, pp. 98-105.
23. C. Kang, X. Wang, Y. Chang, and B. Tseng, "Learning to rank with multi-aspect relevance for vertical search," in *Proceedings of ACM International Conference on Web Search and Data Mining*, 2012, pp. 453-462.
24. Z. Dai, C. Xiong, and J. Callan, "Query-biased partitioning for selective search," in *Proceedings of ACM International Conference on Information and Knowledge Management*, 2016, pp. 1119-1128.
25. Z. Dai, Y. Kim, and J. Callan, "Learning to rank resources," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 837-840.
26. J. Xu and X. Li, "Learning to rank collections," in *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 765-766.
27. T. Wu, X. Liu, and S. Dong, "LTRRS: A learning to rank based algorithm for resource selection in distributed information retrieval," in *Proceedings of the 25th China Conference on Information Retrieval*, 2019, pp. 52-63.
28. A. Kulkarni and J. Callan, "Selective search: Efficient and effective search of large textual collections," *ACM Transactions on Information Systems*, Vol. 33, 2015, pp. 1-33.
29. Y. Kim, J. Callan, J. S. Culpepper, and A. Moffat, "Load-balancing in distributed selective search," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 905-908.
30. D. Hong and L. Si, "Search result diversification in resource selection for federated search," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 613-622.
31. R. Balakrishnan and S. Kambhampati, "Factal: Integrating deep web based on trust and relevance," in *Proceedings of the 20th International World Wide Web Conference*, 2011, pp. 181-184.
32. R. Balakrishnan, S. Kambhampati, and M. Jha, "Assessing relevance and trust of the deep web sources and results based on inter-source agreement," *ACM Transactions on Web*, Vol. 7, 2013, pp. 11-32.
33. T. Rehatsinas, X. L. Dong, and D. Srivastava, "Characterizing and selecting fresh data sources," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 919-930.
34. S. Deng, C. Wan, and X. Liu, "Selection of multimedia data source based on user feedback," in *Proceedings of IEEE International Conference on Management of e-Commerce and e-Government*, 2011, pp. 285-289.
35. X. L. Dong, B. Saha, and D. Srivastava, "Less is more: Selecting sources wisely for integration," in *Proceedings of the VLDB Endowment*, Vol. 6, 2012, pp. 37-48.

36. Y. Lin, X. Hu, and X. Wu, "Quality of information-based source assessment and selection," *Neurocomputing*, Vol. 133, 2014, pp. 95-102.
37. Y. Lin, H. Wang, S. Zhang, J. Li, and H. Gao, "Efficient quality-driven source selection from massive data sources," *Journal of Systems and Software*, Vol. 118, 2016, pp. 221-233.
38. Y. Lin, H. Wang, J. Li, and H. Gao, "Data source selection for information integration in big data era," *CoRR*, 2016, No. abs/1610.09506.
39. B. Catania, G. Guerrini, and B. Yaman, "Context-dependent quality-aware source selection for live queries on linked data," in *Proceedings of the 19th International Conference on Extending Database Technology*, 2016, pp. 716-717.
40. H. Baoli, C. Ling, and T. Xiaoxue, "Knowledge based collection selection for distributed information retrieval," *Information Processing & Management*, Vol. 54, 2018, pp. 116-128.
41. F. Z. Lebib, H. Drias, and H. Mellah, "Selection of information sources using a genetic algorithm," in *Proceedings of the 5th World Conference on Information Systems and Technologies*, 2017, pp. 60-70.
42. J. Lu and J. Callan, "User modeling for full-text federated search in peer-to-peer networks," in *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 332-339.
43. S. Kechid and H. Drias, "Personalised distributed information retrieval-based agents," *International Journal of Intelligent Systems Technologies and Applications*, Vol. 9, 2010, pp. 49-74.
44. E. Abel, J. Keane, N. W. Paton, A. A. A. Fernandes, M. Koehler, N. Konstantinou, J. C. C. Rios, N. A. Azuan, and S. M. Embury, "User driven multi-criteria source selection," *Information Sciences*, Vol. 430, 2018, pp. 179-199.
45. D. Zhou, S. Lawless, X. Wu, W. Zhao, and J. Liu, "Enhanced personalized search using social data," in *Proceedings of ACL Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 700-710.
46. D. Vallet, I. Cantador, and J. M. Jose, "Personalizing web with folksonomy-based user and document profiles," in *Proceedings of the 32nd European Conference on Information Retrieval Research*, 2010, pp. 420-431.
47. M. J. Carman, M. Baillie, and F. Crestani, "Tag data and personalized information retrieval," in *Proceedings of ACM Workshop on Search in Social Media*, 2008, pp. 27-34.
48. D. Zhou, S. Lawless, and V. Wade, "Web search personalization using social data," in *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, 2012, pp. 298-310.
49. M. Aliannejadi and F. Crestani, "Personalized context-aware point of interest recommendation," *ACM Transactions on Information Systems*, Vol. 1, Article No. 1.
50. X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 675-684.

51. J. Trant, “Studying social tagging and folksonomy: A review and framework,” *Journal of Digital Information*, Vol. 10, 2009.
52. M. Gupta, R. Li, Z. Yin, and J. Han, “An overview of social tagging and applications,” *Social Network Data Analytics*, 2011, pp. 447-497.
53. Z. Saoud and S. Kechid, “Integrating social profile to improve the source selection and the result merging process in distributed information retrieval,” *Information Sciences*, Vol. 336, 2016, pp. 115-128.
54. T. L. Griffiths and M. Steyvers, “Finding scientific topics,” in *National Academy of Science*, Vol. 101, 2004, pp. 5228-5235.
55. D. M. Blei, “Probabilistic topic models,” *Communications of the ACM*, Vol. 55, 2012, pp. 77-84.
56. H. Jelodar, Y. Wang, C. Yuan, and X. Feng, “Latent dirichlet allocation (lDA) and topic modeling: models, applications, a survey,” *Multimedia Tools and Applications*, Vol. 78, 2019, pp. 15169-15211.
57. X. Wei and W. B. Croft, “Lda-based document models for ad-hoc retrieval,” in *Proceedings of the 29th European Conference on Information Retrieval Research*, 2006, pp. 178-185.
58. M. Baillie, M. Carman, and F. Crestani, “A multi-collection latent topic model for federated search,” *Information Retrieval*, Vol. 14, 2011, pp. 390-412.
59. M. Harvey, F. Crestani, and M. J. Carman, “Building user profile from topic models for personalised search,” in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 2013, pp. 2309-2314.
60. E. M. Rochd and M. Quafafou, “A topic model-based personalization over time,” in *Proceedings of the 2nd Workshop on User Engagement Optimization at KDD*, 2014.
61. M. J. Carman, F. Crestani, M. Harvey, and M. Baillie, “Towards query log based personalization using topic models,” in *Proceedings of the 19th ACM International Conference on Information & Knowledge Management*, 2010, pp. 1849-1852.
62. R. Giri, H. Choi, K. S. Hoo, and B. D. Rao, “User behavior modeling in a cellular network using latent dirichlet allocation,” in *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning*, 2014, pp. 36-44.
63. F. Zhao, Y. Zhu, H. Jin, and L. T. Yangbc, “personalized hashtag recommendation approach using lda-based topic model in microblog environment,” *Future Generation Computer Systems*, Vol. 65, 2016, pp. 196-206.
64. Y. Jin, R. Li, Y. Cai, Q. Li, A. Daud, and Y. Li, “Semantic grounding of hybridization for tag recommendation,” in *Proceedings of the 11th International Conference on Web-Age Information Management*, 2010, pp. 139-150.
65. C. Chelmis and V. K. Prasanna, “Social link prediction in online social tagging systems,” *ACM Transactions on Information Systems*, Vol. 31, 2013, Article 20.
66. Y. Liao, W. Lam, L. Bing, and X. Shen, “Joint modeling of participant influence and latent topics for recommendation in event-based social networks,” *ACM Transactions on Information Systems*, Vol. 36, 2018, pp. 1-31.

67. G. Salton, E. M. Voorhees, and E. A. Fox, "A comparison of two methods for boolean query relevance feedback," Technical Report No. 564, Department of Computer Science, Cornell University, 1983.
68. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., NY, 1986.
69. A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles, "The impact of database selection on distributed searching," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 232-239.
70. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "INDRI: A language model based search engine for complex queries," in *Proceedings of International Conference on Intelligent Analysis*, Vol. 2, 2005, pp. 2-6.
71. W. B. Croft and J. Callan, "Lemur project," https://www.lemurproject.org/, 2000.
72. X.-H. Phan and C.-T. Nguyen, *JGibbLDA*, 2008.
73. F. G. Lobo, D. E. Goldberg, and M. Pelikan, "Time complexity of genetic algorithms on exponentially scaled problems," in *Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation*, 2000, pp. 151-158.
74. P. S. Oliveto, J. He, and X. Yao, "Time complexity of evolutionary algorithms for combinatorial optimization: A decade of results," *International Journal of Automation and Computing*, Vol. 4, 2007, pp. 281-293.

Fatma Zohra Lebib is a Researcher at Research Center on Scientific and Technical Information (CERIST). She is Ph.D. at the University of Science and Technology Houari Bomediene (USTHB). Her research interests focus on distributed information retrieval,web semantic, adaptability and social context.



Hakima Mellah is a researcher at Research Center on Scientific and Technical Information (CERIST). Her researches concern interacting distributed and agile information systems based on multi-agents systems.

Abdelkrim Meziane is a Researcher and a responsible at Research Center on Scientific and Technical Information (CERIST). His interest research areas include image analysis, hospital information systems, and medical imaging.