A Knowledge-based Model for Semantic Oriented Contextual Advertising

Mohammed Maree[1,*], Rami Hodrob[1], Mohammed Belkhatir[2] and Saadat M. Alhashmi[3]
[1] Department of Information Technology, Arab American University
Jenin, Palestine
[e-mail: {mohammed.maree, rami.hodrob}@aaup.edu]
[2] Campus de la Doua, University of Lyon
Lyon, France
[e-mail: mohammed.belkhatir@univ-lyon1.fr]
[3] Department of Management Information Systems, University of Sharjah
Sharjah, UAE
[e-mail: salhashmi@sharjah.ac.ae]
*Corresponding author: Mohammed Maree

Received September 15, 2019; revised February 22, 2020; accepted March 5, 2020; published May 31, 2020

Abstract

Proper and precise embedding of commercial ads within Webpages requires Ad-hoc analysis and understanding of their content. By the successful implementation of this step, both publishers and advertisers gain mutual benefits through increasing their revenues on the one hand, and improving user experience on the other. In this research work, we propose a novel multi-level context-based ads serving approach through which ads will be served at generic publisher websites based on their contextual relevance. In the proposed approach, knowledge encoded in domain-specific and generic semantic repositories is exploited in order to analyze and segment Webpages into sets of contextually-relevant segments. Semantically-enhanced indexes are also constructed to index ads based on their textual descriptions provided by advertisers. A modified cosine similarity matching algorithm is employed to embed each ad from the Ads repository into one or more contextually-relevant segments. In order to validate our proposal, we have implemented a prototype of an ad serving system with two datasets that consist of (11429 ads and 93 documents) and (11000 documents and 15 ads), respectively. To demonstrate the effectiveness of the proposed techniques, we experimentally tested the proposed method and compared the produced results against five baseline metrics that can be used in the context of ad serving systems. In addition, we compared the results produced by our system with other state-of-the-art models. Findings demonstrate that the accuracy of conventional ad matching techniques has improved by exploiting the proposed semantically-enhanced context-based ad serving model.

Keywords: Ad matching, experimental evaluation, contextual advertising, semantic resources, relevance judgments

1. Introduction

Contextual Advertising (CA) refers to the proper embedding of commercial ads (henceforth referred to as ads) within the content of generic publisher Webpages, to provide a better user experience on the one hand, and to increase the revenues of both publishers and advertisers on the other [1, 2]. Current Ad serving systems either rely on traditional keyword overlap techniques, or employ semantic resources to accomplish the matching task between ads and Webpages. As far as keyword-based matching techniques are concerned, it is clear that systems that employ such techniques still suffer from low precision ratio i.e. a large fraction of the systems' results are irrelevant [3]. On the other hand, newer semantics-based approaches are penalized by limitations of the exploited semantic resources, namely semantic knowledge incompleteness, limited domain coverage and time consumption problems [4, 5]. In addition, current contextual advertising approaches treat each Webpage as a whole segment, ignoring that fact that it can comprise several segments that may cover more than one topic or domain. Acknowledging these drawbacks, we propose an integrated ad selection and serving approach that employs knowledge encoded in publicly available domain-specific and generic knowledge bases, in addition to a modified cosine similarity metric to cooperatively capture the semantic aspects that are latent in the content of publisher Webpages, as well as in the ads' textual descriptions. In this context, we use the knowledge bases to construct semantic indexes for the ads and their corresponding Webpages. Then, we employ the modified cosine similarity technique to find matches between the constructed indexes and embed each ad within its semantically-relevant segment of each Webpage. The used knowledge resources are assisted by other statistical (term co-occurrence) and NLP techniques (stop word removal, Named Entity Recognition (NER) and Part of Speech tagging (POS)) to effectively address contextual matching and retrieve relevant ads for the targeted audience. Accordingly, we summarize the goals of this research into the following points:

- The integration of semantic knowledge and statistical-based semantic similarity measures within a unique contextual advertising scenario.
- The specification and design of a precision-oriented ad serving system considering the semantics and context of both ads and their corresponding contextual segments within publisher Webpages.

To achieve the aforementioned goals, we carry on the research work on two important fields: 1) semantics-based content analysis, and 2) contextual ad matching. These two fields are strongly related in our proposed approach as they rely on the development and use of formal advertising concept standards and metrics [6-8]. To validate the effectiveness of the proposed system, we define an interface that facilitates human interaction and evaluate our precision-oriented context-based ad serving system using state-of-the-art indicators for advertising evaluation, namely precision.

The remainder of this article is organized as follows. In Section 2, we present a background on the evolution of contextual advertising and highlight the strengths and limitations that are associated with existing approaches. Section 3 introduces the formal definitions and discusses the problem formulation. Section 4 provides a general overview of the architecture of our proposed system. Section 5 presents the experimental evaluation of the proposed system. In this section, we compare between the results produced by our system against other state-of-the-art baseline metrics. Finally, in Section 6, we present the conclusions and discuss the future work.

2. Related Work

The goal of contextual advertising is to help advertisers reach consumers with a strong preference for their products [2, 7, 9]. In this context, contextual advertising systems attempt to automatically accomplish this task through the placement of the most relevant commercial textual ads within the content of generic publisher Webpages [10, 11]. As stated by Kumar et al. [11], contextual advertising employs automated techniques that display advertisers' ads that are relevant to the user's interests as well as publisher website's content.

Several approaches have been proposed to realize contextual advertising in practical and real-world application domains [7-9, 12-27]. The authors of [1], [7] and [13] have exploited knowledge encoded in Wikipedia to interpret the textual content of ads. The goal of these systems is to overcome problems such as homonymy and polysemy of ambiguous ad terms, as well as low intersection between the keywords of Webpages and their corresponding ads. In this context, prior to matching ads to their relevant Webpages, Wikipedia is utilized as an intermediate reference model for enriching the representation of the documents with semantically-related terms. However, these approaches suffer from a major problem associated with the exploitation of Wikipedia as a reference model for discovering the hidden semantic dimensions in the content of documents (ad texts and Webpage text fragments). This is due to the fact that Wikipedia is characterized by limited concepts coverage and lacks a formal and explicit definition of the semantic relations that may hold between the concepts extracted from the documents. In addition, mapping concepts to a huge number of Wikipedia articles may result in a serious time consumption problem, making it difficult to utilize it in practical settings. On the other hand, the proposed approaches attempt to match between ads and Webpages, ignoring the fact that the content of a Webpage may cover multiple topics (represented by textual segments that describe different domains). Ignoring this important aspect about Webpages may cause ad matching techniques to i) embed a single ad under various topics within the same page, ii) embed multiple ads inside the content of a single page, or iii) place a single ad within multiple pages. Hence, the precision of ad matching techniques degrades due to these reasons.

Other researchers propose to exploit text summarization and probabilistic techniques to improve the matching process between ads and their corresponding Webpages [9, 15, 25]. Using text summarization techniques, text fragments of Webpages are summarized into short textual descriptions to minimize the matching space and reduce semantic dimensionality problems [15, 25]. However, these approaches rely on using text processing algorithms that ignore the semantic aspects of the extracted texts and therefore, result in producing low-precision matching results. As far as probabilistic models are concerned, their performance is limited due to their dependence on other resources such as existing search engines as demonstrated in the context of the work proposed in [9]. As stated by the authors, using search engines such as Google and Baidu to measure the relatedness between terms is a limiting factor to their proposed approach since if the search engine performs poor for some terms, their proposed method may produce poor results as well. In a similar line of research, Zheng et al. [28] utilized a taxonomy tree based contextual advertising model, where they created a training set to populate the taxonomy with concepts to classify both advertisements and web pages based on their taxonomic relations in the tree. The authors argued that the developed taxonomic structure has proved to work well in Chinese documents. However, the proposed approach still requires a huge manually constructed training set to populate the taxonomy with new concepts. In addition, the accuracy of the proposed approach relies on the quality of the produced taxonomy, which is a similar problem to the problem of probabilistic models that we have discussed in this section. Moreover, this model can be applied on a

limited subset of Webpages, i.e. they can be applied only to limited languages. Due to the limitations discussed in the related works, we attempt to address and overcome the problems that are associated with current techniques through the incorporation of semantic knowledge and statistical-based semantic similarity measures within a unique contextual advertising scenario. In this context, we aim to identify and enrich the semantic matching procedure with additional concepts that are not recognized by current semantic repositories. The integration of semantic resources in our proposed model forms a primary knowledge source in our context. Additionally, unlike probabilistic models that are limited due to their dependence on other external resources that may produce poor results, we exploit semantic resources that are characterized by a manual confirmed accuracy of 95% as reported in [29].

3. Problem Formulation

In this section, we define a set of notations which will be used throughout this work. Let D be a set of ads database which contains N_{ad} ads, represented by $D = \{a_j\}_{j=1}^{N_{ad}}$. Let A denote an advertisement which is composed of a set of words $\{w_i \mid i \in [1,N]\}$. Let W_p be a targeted Webpage used to match candidate ads. W_p is composed of a set of sentences $\{s_j \mid j \in [1,N]\}$. Similar to A each s_i is composed of a set of words $\{w_j \mid j \in [1,N]\}$. One approach to finding the similarity between $A_i \in D$ and W_p is to find the cosine similarity between the words $\{w_i \mid i \in [1,N]\}$ in A_i and the words $\{w_j \mid j \in [1,N]\}$ in $s_i \in W_p$. The cosine similarity metric a.k.a. vector space model, employs the $tf-idf$ weighting scheme [30] for deriving the vectors of $A_i \in D$ and W_p. The original form of the $tf-idf$ weighting scheme is normally employed to assign a term t a weight w in a document d as shown in Equation 1:

$$tf-idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

Where $tf_{t,d}$ is the number of occurrences for term t in d and is assigned using the below equation:

$$tf_{t,d} = \begin{cases} 1+\log_{10} tf_{t,d} & if \quad tf_{t,d} > 0 \\ 0, & Otherwise \end{cases} \tag{2}$$

Where the idf_t is the inverse document frequency of the term t and is assigned as follows:

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right) \tag{3}$$

The cosine similarity metric deals with $A_i \in D$ and W_p as vectors, and to find the similarity between a given pair A_i and W_p, the following formula is used:

$$\cos ine(A_i, W_p) = \frac{\vec{V}(A_i).\vec{V}(W_p)}{\left|\vec{V}(A_i)\right|\left|\vec{V}(W_p)\right|} \tag{4}$$

As shown in Equation 4, the employed similarity metric attempts to find a relevance score between a given ad A_i and a given Webpage W_p based on their dot product. However, this

metric ignores the hidden semantic dimensions in the text of the ads, as well as in the text of Webpages. The fact that some terms (either in A_i or in W_p) are synonymous or are semantically related to other terms is not incorporated in this model. To overcome this issue, we propose exploiting additional semantic resources that can be employed to discover the latent semantic relations in texts and accordingly lead to a more precise matching procedure. In this context, a semantic resource can be defined as follows:

Definition 1: Semantic Resource:

A semantic resource SR is *quintuple, SR*:= (*C, P, I, V, A)* where:

- *C* represents the set of concepts that are defined in *SR*. The hierarchical relationship between concepts of the set *C* is a pair (*C*, ≤), where ≤ is an order relation on *C* x *C*. We call ≤ the sub-concept relation.
- *P* represents the set of properties defined over *C*.
- *I* is the set of individuals also called instances of the concepts in *SR*.
- *V* is the set of values defined over *P*.
- *A* is the set of axioms in *SR*.

For a given pair A_i and W_p, we first find the $tf-idf$ weights for the terms of both documents. Then, terms with high $tf-idf$ weights $H_w = \{h_j \mid j \in [1, N]\}$ will be redirected towards a dedicated semantic resource for finding their synonyms. In addition to the synonymy relation, we obtain the first hypernym of each term as we believe that it can assist in finding more pages that are suitable for embedding the ads within their corresponding contextual segments. For instance, assume the term "car" represents w_i in a given $A_i \in D$ and that "car" $\in$ H_w. Synonymous terms such as "auto" and "automobile", as well as the hypernyms such as "motor vehicle" and "automotive vehicle" are obtained when using WrodNet [31] semantic resource. We would like to point out that although WordNet has been widely used in similar research works, we believe that it will not be sufficient in our work. This is because it has a limited domain coverage and also it is not a multilingual resource. There are other versions of WordNet in other languages, but they are not integrated with the original English version. Therefore, in our work, we exploit additional resources such as YAGO3 [29] semantic resource. YAGO3 is an extension of the YAGO knowledge base that is derived from Wikipedia WordNet and GeoNames. It combines the information from Wikipedia in multiple languages. YAGO3 is an enlarged version of YAGO (which has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities) with 1 million new entities and 7 million new facts.

Definition 2: Semantically-enhanced $tf-idf$ Weighting:

As we have discussed in this section, the $tf-idf$ weighting scheme ignores the hidden semantic dimensions in the text of the ads, as well as the text of Webpages. To overcome this issue, we propose exploiting the semantic resource SR in order to enhance the precision of the conventional $tf-idf$ weighting scheme. In this context, terms in H_w are submitted to SR to find their synonyms, as well as the first hypernym (by moving one level up in the hierarchical structure of the semantic resource) of each $\{h_j \mid j \in [1, N]\}$. Accordingly, we use the below modified version of the $tf-idf$ weighting scheme:

$$freq(h_i) = \begin{cases} 1+\log_{10} tf(h_i) \times \mathrm{Sy}(h_i) \times \mathrm{Hy}(h_i) & if \quad tf_{t,d} > 0 \\ 0, & Otherwise \end{cases} \tag{5}$$

Where,

- $freq(h_i)$: is the total number of occurrences of each h_i .
- $Sy(h_i)$: is the number of synonymous terms of each h_i . We may find one or more synonyms for each h_i . For instance, both terms “auto” and “automobile” are synonymous of the term “car”.
- $Hy(h_i)$: is the number of hypernyms of each h_i . By moving one level up in the hierarchical structure of the semantic resource, we may find one or more hypernyms for each h_i . For instance, both terms “motor vehicle” and “automotive vehicle” are synonymous parents of the term “car”.

We also, use the below modified version of the idf_t formula:

$$Midf_t = \log_{10}\left(\frac{N}{Modified\_df_t}\right) \tag{6}$$

Where,

- N: is total number of documents (Ads or Webpages) in the collection.
- $Modified\_df_t = df(h_i) \times \mathrm{Sy}(h_i) \times \mathrm{Hy}(h_i)$

Definition 3: Ad Zone Weighting:

In our approach, an ad is divided into zones wherein each zone is given a weight. Let $w_1,...,w_l \in [0,1]$ such that $\sum_{i=1}^{l} w_i = 1$ for $1 \le i \le l$, and let s_T , s_B and s_K represent the title, body and keyword zones of each ad. Then, the weighted zone score is defined as:

$$AdZone = w_1.s_T + w_2.s_B + w_3.s_K \tag{7}$$

In the proposed approach, we give weights 0.5, 0.2, 0.3 to s_T , s_B and s_K respectively. The reason behind giving these weights is because we believe that both the Title of an ad and its associated keywords have greater contribution to the semantic interpretation of the ad than the text of its body. It is important to point out that, unlike conventional zone scoring techniques (where each zone of the ad contributes a boolean value i.e. 1 when a term t_i appears in a zone, and 0 if it does not occur in the zone) the contribution of the three zones s_T , s_B and s_K is computed as follows:

- $s_T =$ $freq(t_i)$ in $s_T * w_1$
- $s_B =$ $freq(t_i)$ in $s_T * w_2$
- $s_K =$ $freq(t_i)$ in $s_K * w_3$

Accordingly, when matching ads to segments of Webpages, the ad Matching algorithm exploits Equation 7 for assigning relevance scores for placing ads in their appropriate locations.

Definition 4: Ad Matching:

The ad Matching algorithm takes a given ad A_i from D and a given s_i from W_p as input and finds the relevance scores between the words $\{w_i \mid i \in [1, N]\}$ in A_i and the words $\{w_j \mid j \in [1, N]\}$ in $s_i \in W_p$ based on employing the modified version of the $tf - idf$ weighting scheme:

$freq(w_j)$ x $Midf_{w_j}$. The cosine similarity metric in this context deals with the vectors obtained from $A_i \in D$ and $S_i \in W_p$, and is computed using the following formula:

$$\cos ine(A_i, S_i) = \frac{\vec{V}(A_i).\vec{V}(S_i)}{\left|\vec{V}(A_i)\right|\left|\vec{V}(S_i)\right|} \tag{8}$$

The details of this step are illustrated in Algorithm 1 below.

Algorithm 1. Ad Matching Algorithm

Input: AdsList and Webpage Segments
Output: relevance scores between each ad A_i and each segment S_i of a given Webpage *Wp*

```
1: float Scores[N]=[0] // is an array with a score entry for each Ai, initialized to zero.
2: constant w[l] //Assume w[l] is initialized to the respective zone weights
3: Ads← getList(Ads)
4: Segments← getList(Wp)
5: result ← ⟨ ⟩;
6: for each Ai in Ads
7: for each Si in Wp
8: Scores [ Ai ]= cosine(Ai, Si)
9: end for
10: end for
11: return Top K components of Scores[] // K= 5 in our work.
```

4. Detailed Characterization of the Proposed Solution

There are three main actors (Advertisers, ad Networks, Publishers) in the contextual advertising domain. Advertisers create ad campaigns and provide details about their ads, such as titles, descriptions, icons or images, bid terms, as well as keywords to be involved in the matching process. Publishers on the other hand offer a space for embedding ads in their Webpages. Normally, there is no direct relation between publishers and advertisers as ad networks usually take care of connecting both parties. In this context, advertisers pay for ad networks which guarantees routing the ads to the publishers' websites. Publishers gain profit from displaying advertiser ads, and ad Networks share part of that profit. In the context of our work, our proposed ad matching procedure can be employed by ad networks to precisely select the most relevant Webpages for placing ads. Additionally, the exploited knowledge bases can be utilized to assist advertises in selecting the most appropriate keywords (in addition to their semantically related terms) when describing an ad. In **Fig. 1** below, we provide an overview of the basic flow of our proposed ad matching procedure.

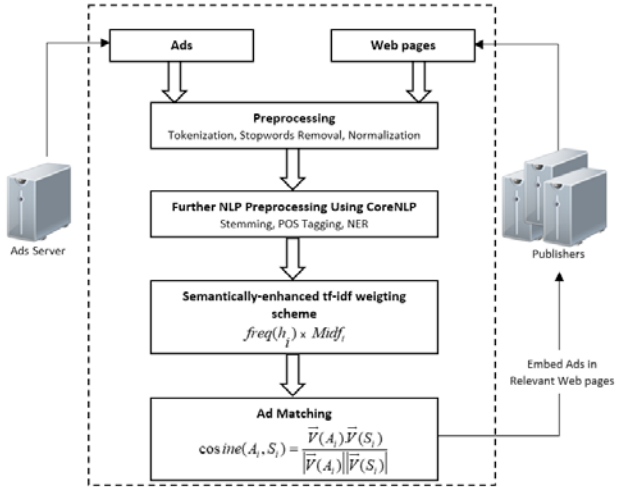

Fig. 1. Basic flow of the Ad matching procedure

As shown in **Fig. 1**, the content of both ads and Webpages is pre-processed through applying tokenization, stopwords removal, and normalization steps. We would like to point out that we employ the traditional tf-idf weighting scheme at the corpus level in order to remove the set of stopwords SW which is obtained based on a threshold value v using Equation 9 below:

$$SW = \{s_w \in SW \mid tf - idf(s_w) \leq v\} \tag{9}$$

The content of the ads and Webpages will be further preprocessed using the Stanford CoreNLP:

- Part-of-Speech Tagging: each token is assigned to its part of speech category such as noun, verb, adjective, etc. To accomplish this task, we employ the Stanford CoreNLP POSTagger.
- Stemming: some tokens can be brought under the same category (common base form) after the removal inflectional forms such as derivational affixes.
- Named Entity Recognition (NER): during this step, each token is assigned a category based on a set of pre-defined categories such as person, organization, and location.

Next, we employ knowledge encoded in the semantic resources to semantically enhance the extracted terms. In this context, we find the synonyms, as well as the first hypernym for each term. As highlighted in section 3, we utilize the modified version of the cosine similarity metric for computing the similarity scores between each ad and its corresponding segment from each Webpage.

5. Experimental Setup and Evaluation

In this section, we introduce the experimental evaluation steps that we have followed in order to evaluate the quality of the proposed contextual-advertising approach. We would like to point out that solutions have been implemented using Java programming language on a PC with core i7 CPU (2.1GHz) and (16 GB) RAM with Windows 10 operating system.

We have used two different datasets in order to evaluate the quality of the proposed model, and compare our results against other related works to demonstrate the effectiveness of the propose semantically-enhanced techniques. We provide details of both dataset as follows:

1. The first dataset is a publicly-available dataset that has been extensively used by several semantic matching approaches over the past years. This dataset is composed of 93

documents that we used to represent Webpages and a collection of 11429 textual segments that we used to represent our ads collection. It is important to point out that instead of utilizing web crawling techniques to collect Webpages in the same manner as performed in [9], we decided to use the documents in this dataset since they are available online and can be utilized for the same purpose by other researches in the field. In addition, the language used to express documents in this dataset is English, and since our exploited ontologies and semantic similarity techniques are applied on English words, we were not able to reuse datasets in languages other than this language. More importantly, in order to avoid bias and maintain the reproducibility of the produced results, relevance judgements between both collections are also provided by the authors of the dataset. The availability of these judgments enabled us to compare between the results produced by our system with those assigned by subject matter experts. We summarize the reasons behind our decision to use this dataset as follows:

- First, we aim to experimentally validate our proposal and demonstrate the accuracy of utilizing the proposed techniques in enhancing the similarity results between ads and their corresponding web documents. The results of employing the proposed techniques using this dataset are further detailed in the next section.
- Second, to our knowledge there is no publicly-available advertising gold-standard or ground truth that comprises ads, Webpages and the relevance scores between each ad and its corresponding Webpages/s in the English language in the same manner as provided in this dataset. Therefore, instead of constructing the dataset from scratch, we decided to use a dataset that can serve as our gold-standard through which we can evaluate our system's results and compare them with other models.
- Third, as the authors of the dataset offer the relevance judgements, this assists in enabling other researchers in this domain to use and re-use the dataset - including our results - to guarantee the reproducibility of the experiments and to be further employed in future research studies in the domain.

We would like to point out that in order to build initial indexes for the documents in the dataset, we have used Indri [32]. **Fig. 2** depicts the UI of Indri after running the Build Index function.

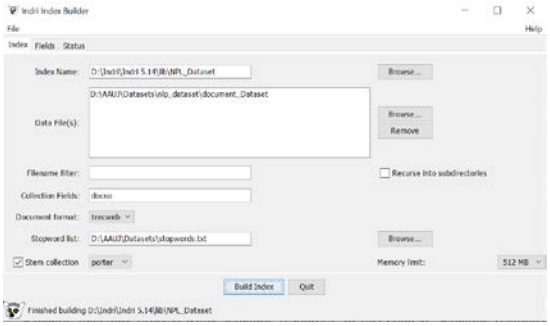

Fig. 2. A screenshot from Indri's UI after running the Build Index function

2. We used the second dataset to perform experiments with the aim of comparing the techniques utilized in the proposed model against other related techniques that have been proposed to address contextual-based advertising. To assess the quality of the proposed techniques, we used the BankSearch dataset [33], that was constructed using the Open Directory Project and Yahoo! Categories in the same manner as performed by [34]. The dataset consists of 11000 Webpages that are manually classified under 11 different classes as depicted in **Fig. 3**. It is important to note that the 11 selected classes are the leaves of the

taxonomy, together with the class *Sport*. This class is included among other leaf nodes as it contains documents from sites that were classified as *Sport*, but not *Soccer* or *Motor Sport*. In addition to the Webpages, we have collected 15 ads to calculate the precision of the proposed ad matching model and compare the results against state-of-the-art matching methods.

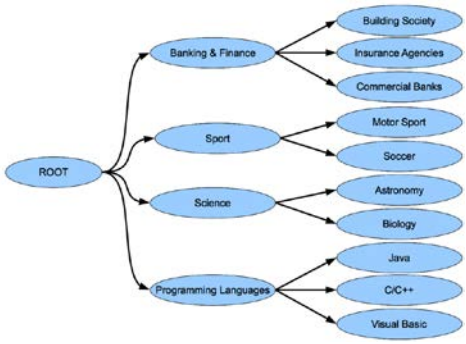

Fig. 3. Class hierarchy of BankSearch dataset as reported in [34]

To pre-process and filter out the documents in the BankSearch dataset, we used RapidMiner[1] software package. **Fig. 2** shows the main NLP steps that we have defined and applied on the documents in the dataset.

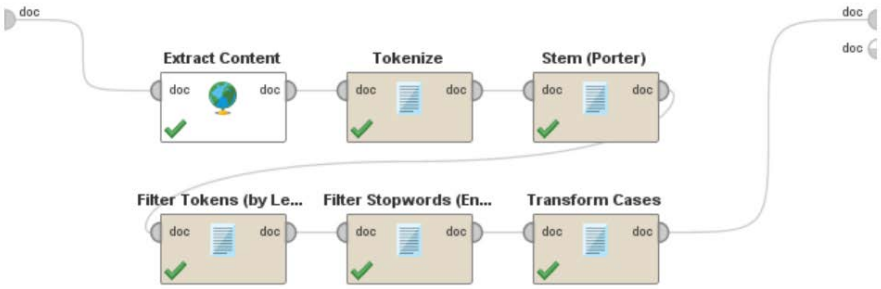

Fig. 4. Main NLP steps used to process documents in BankSearch dataset

5.1 Experiments Using the Proposed Model

In these experiments, we aim to demonstrate the impact of enhancing the Ad-to-Webpage semantic matching process using semantic resources. In this context, semantic relations as well as additional semantically-relevant entities that are extracted from the employed ontologies are utilized for enhancing the baseline edit-distance similarity measures. To validate the effectiveness of our proposal, we start with the first dataset along with five baseline results that are depicted in **Fig. 5**. To obtain these results, we have used five similarity measures in order to calculate the similarity between the text segments that are used to represent the ads and their counterparts that are used to represent Webpages. The used measures are: Cosine Similarity, Jaccard Similarity, JaroWinkler Similarity, Normalized Levenshtein and SorensenDice Similarity. For more details on these measures, please refer to [35] , where the authors provide a summary of a variety of edit distance measures including the ones that we use in our experiments. For demonstration purposes, we have selected the first

[1] https://rapidminer.com/

seven Webpages and their corresponding ad contextual segments. We have selected 10 ad segments per each Webpage.

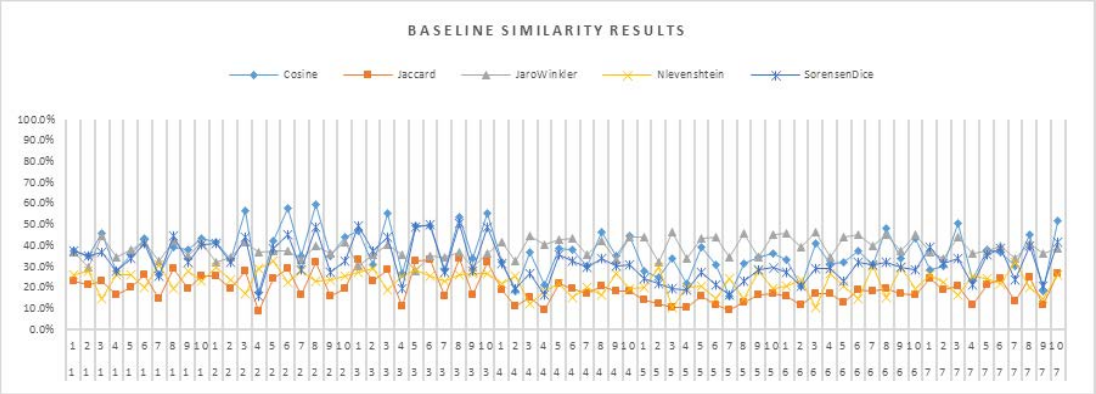

Fig. 5. Baseline similarity results using the five edit distance techniques

As depicted in **Fig. 5**, the produced similarity scores using the five edit distance measures are characterized by their low quality. It is important to point out that the employed techniques are shingle-based, and we have set the number of shingles to be equal to two since we have empirically discovered that this number of shingles leads to producing more accurate results compared with other numbers of shingles. The reason behind the low accuracy results is because we ignore all semantic and taxonomic relations that exist between the terms that are used to describe the textual content of the used documents in the dataset. As such, the used techniques relied only on the actual terms without any sort of semantic or taxonomic enrichment. In the next experiment, we attempted to enrich the indexes of both ads and Webpages with synonyms that are extracted from the employed ontologies. Our goal in this context is to investigate the impact of this enrichment procedure on the quality of the produced results. We present the results of this step in **Fig. 6**. As we can see in this figure, there is an improvement on the overall precision results for the five similarity techniques. However, the level of improvement in this context is not constantly growing for the five techniques. For some of the used similarity techniques, such as the Cosine similarity, introducing synonyms has resulted in a notable improvement on the quality of the matching step. The same also applies to the rest of the employed techniques but with different levels of enhancement.

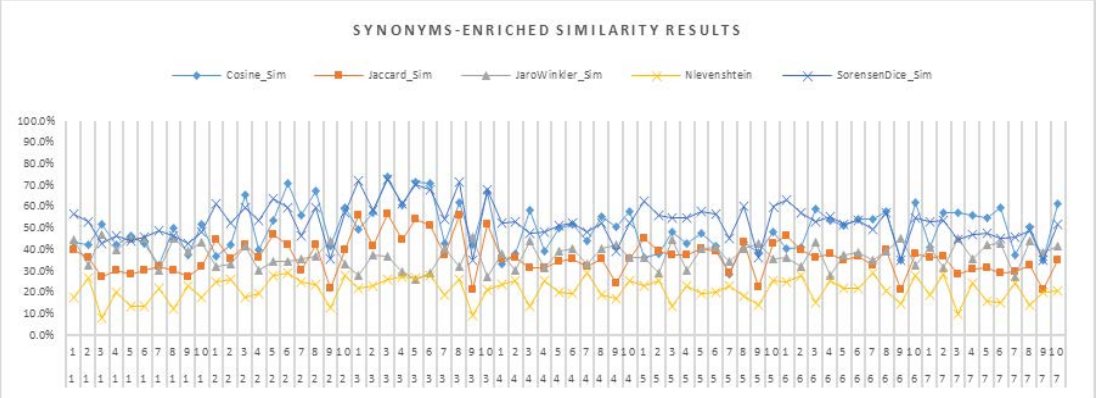

Fig. 6. Similarity results after enriching the indexes with synonyms extracted from the used ontologies

After this step, we further expanded the constructed semantic indexes with additional semantically-related terms such as hypernyms, hyponyms and meronyms as well as using other conceptual-related terms using the employed ontologies. **Fig. 7** shows the results of carrying out this step.

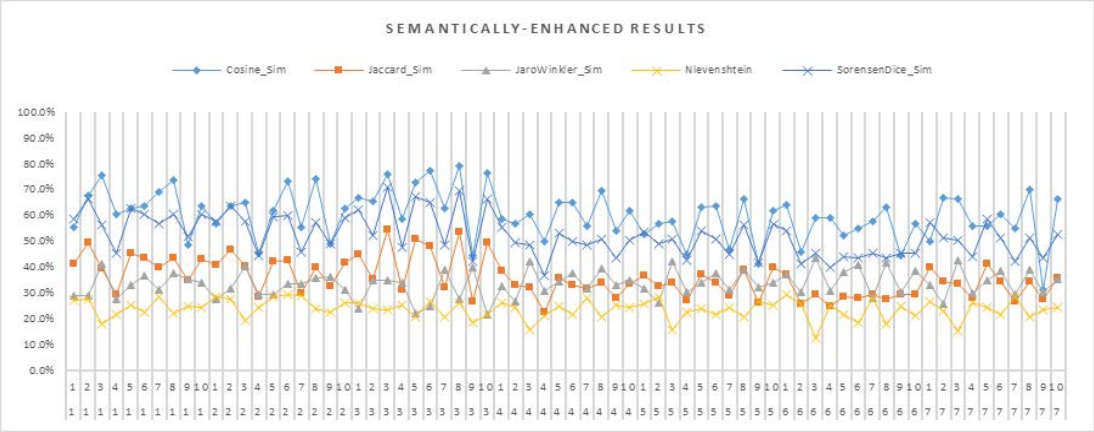

Fig. 7. Similarity results after enriching the semantic indexes with semantically-related terms extracted from the used ontologies

As depicted in **Fig. 7**, a significant level of improvement upon the overall quality of the semantic matching process has been achieved. We demonstrate, using the produced results, that this improvement supports our argument that semantic resources can play a significant role in improving the quality of baseline edit-distance measures when used for matching ads to their corresponding Webpages. In the next figures, we demonstrate the levels of improvement that have been achieved when semantically enhancing each of the five similarity measures. We particularly compare between the quality of the produced results when using the baseline measures against using the synonyms-enriched and semantically-enriched similarity techniques. Our aim of this comparative analysis is to describe the level of improvement obtained on each technique individually, and to draw our recommendations on which technique/s to use in the context of our domain. As depicted by the below figures, the proposed model has played an essential role in improving the quality of the ad matching process.

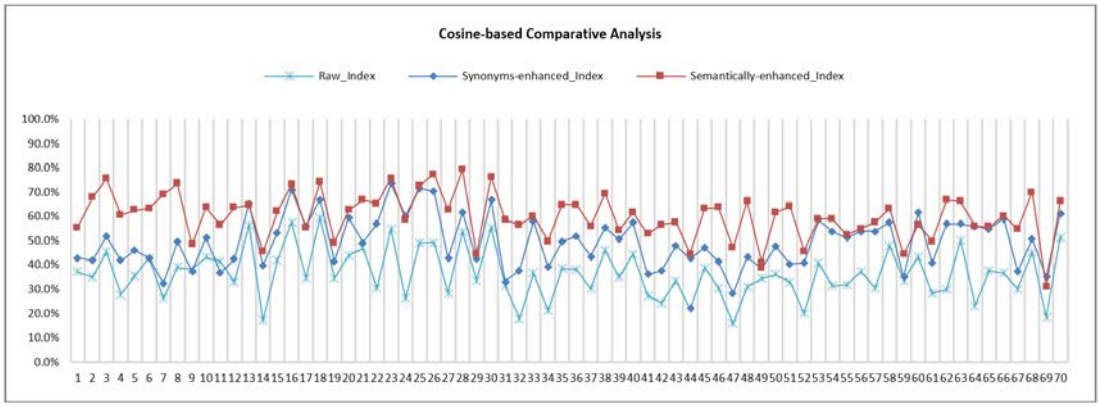

Fig. 8. Baseline against semantically-enhanced Cosine-based similarity results

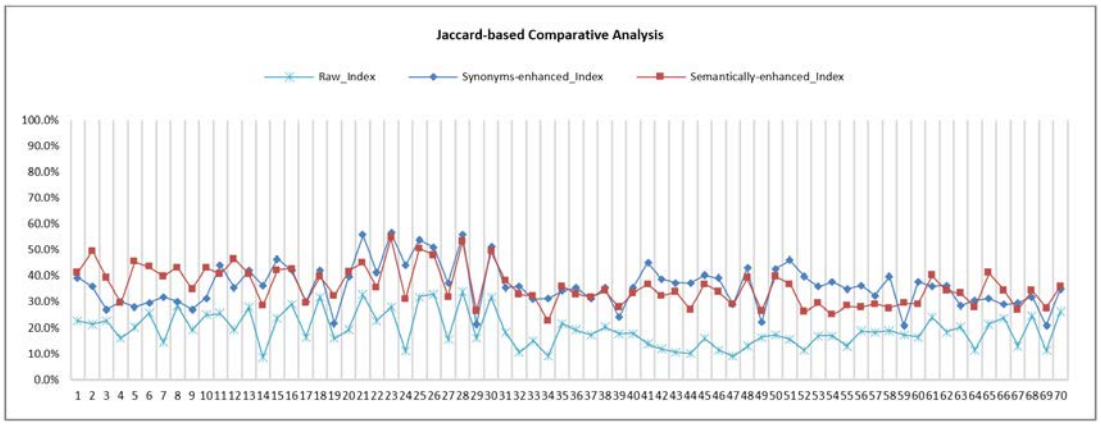

Fig. 9. Baseline against semantically-enhanced Jaccard-based similarity results

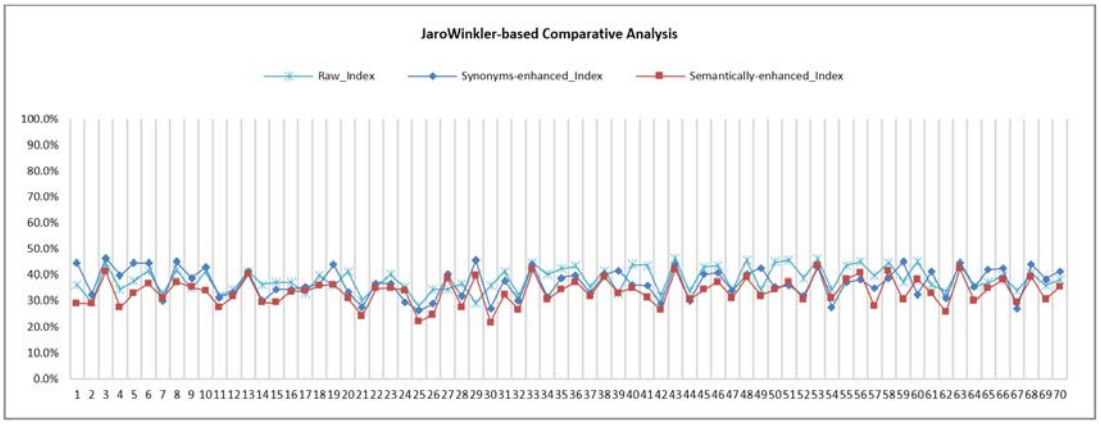

Fig. 10. Baseline against semantically-enhanced JaroWinkler-based similarity results

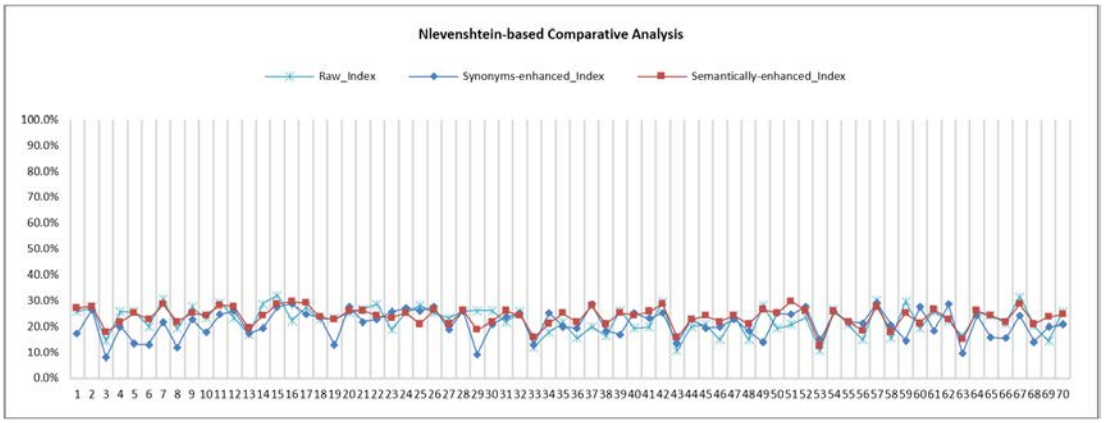

Fig. 11. Baseline against semantically-enhanced Nlevenshtein-based similarity results

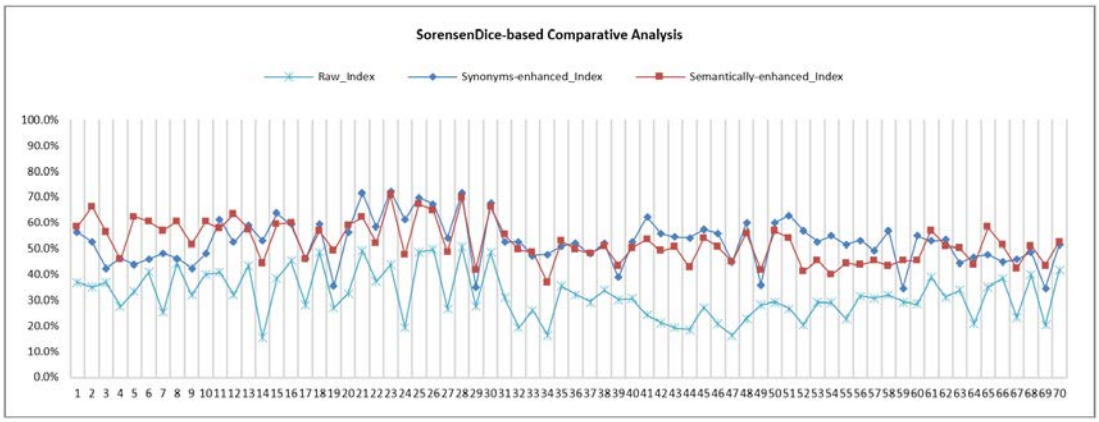

Fig. 12. Baseline against semantically-enhanced SorensenDice-based similarity results

As shown in **Fig. 8-12**, the overall effectiveness of the five measures has improved when applying the proposed semantically-enhanced similarity model. We can also see that the achieved levels of improvement varied from one technique to another. For instance, as depicted in **Figs 8 and 12**, a significant level of improvement has been achieved when extending the baseline metric with semantically-related entities that were not explicitly defined in the textual representations of both ads and Webpages. Indeed, there are improvements on the quality of the rest of the techniques, however, we notice that the highest level was achieved for the Cosine similarity and SorensenDice measures. Accordingly, we would like to point out that updating these measures by incorporating the semantic model can lead to more accurate matching results, and accordingly both publishers and advertisers can gain mutual benefits when utilizing the proposed model. Another important component that can be also utilized using the proposed model, is the possibility of using it for suggesting candidate terms for expressing and describing the desired content for each ad. In this context and unlike conventional approaches that are currently employed by Ad networks, advertisers can better describe their ads with terms that are both contextually and semantically relevant.

In the next experiments, we compare our results using two state-of-the-art models that have been used for testing the effectiveness of contextual ad matching techniques [36]. These are the 1) Bag Of Words (BOW) model alone and 2) BOW with class features (CF). Our aim in this context is to demonstrate the impact of incorporating semantic information on improving the quality of both models, and also compare this with the results produced by the system proposed in [37]. **Table 1** shows the precision of the employed techniques using the BankSearch dataset. We would like to point out that we have compared the results produced by our model (*raw results*: are same as those produced by the BOW model, and *semantically-enhanced results*: to compare with the BOW enriched with CF features, where synonymous concepts and their hypernyms are our features in this model) with the MT model – Paragraph with most title-words – that is used in [37]. Our decision to compare with the MT model was due to the fact that we have considered the same features of each document among the BankSearch dataset.

Table 1. Results of comparing contextual ad matching techniques using the BankSearch dataset

Model	Accuracy Results
BOW – *Raw results*	0.550
BOW with CF - *semantically-enhanced*	0.658
MT Model [37]	0.581

As shown in **Table 1**, for the first model (BOW), both our system and the MT model proposed in [37] have produced comparable results. This is mainly because, in both models, we are treating documents as just a bag of words, ignoring any semantic or taxonomic relation that may exist between the words in ads and their corresponding Webpages. Therefore, only relying on this model will result that are characterized by their poor accuracy. However, as we can see in **Table 1**, for the *semantically-enhanced* model, our system produced more accurate results when we compare it with the MT model. We believe that the incorporation of semantically-enhanced techniques has assisted in better identifying the semantic orientation of ads and their corresponding Webpages. As such, ads were embedded in a more accurate manner than the way they were placed using the BOW or MT models. Again, using the BankSearch dataset, our proposed model has proved to be effective and produced promising accuracy results when compared to conventional contextual matching models.

5. Conclusion

In this paper, we have proposed enhancing conventional edit-distance similarity measures through the exploitation of knowledge encoded in semantic repositories that are available for public usage. The proposed semantically-enhanced model has been employed to address one of the pressing issues; that is precise matching and placement of commercial ads within the context of their corresponding Webpages. A review and analysis of existing ad matching models according to a set of evaluation criteria is introduced in this article. Additionally, we have presented our proposed solution and discussed its components, with an emphasis on the implementation steps that are required to fully develop and deploy the proposed model in a real-world scenario. Unlike conventional systems, the proposed model employs knowledge encoded in semantic resources to improve the quality of the ad matching process and also to suggest candidate terms that can assist advertisers better describe their advertisements. We experimentally demonstrated the impact of using the proposed model on the quality of five baseline similarity techniques. We have also highlighted the levels of improvement achieved for each of the studied techniques and provided our recommendations on using two popular measures (cosine similarity and sorensendice) for accomplishing the ad matching process. The conducted experiments were carried out using two real-world datasets that comprise textual segments that are used to represent ads and Webpages. The produced results showed promising precision improvements and proved the effectiveness of the employed techniques in assigning relevance scores between candidate advertisements against their Webpage counterparts. Nevertheless, it is important to point out that there are still a number of limitations in the current version of the proposed model. Among these limitations is the complexity of the matching algorithm due to the utilization of several external resources that consist of the ontologies and other external libraries that are required for text processing. The complexity of the algorithm will also increase with each addition of a new sematic resource. To overcome this limitation, we plan to construct a single integrated semantic resource that comprises several domain-specific and generic ontologies that can aid during the matching process. In the future work, we plan to construct additional datasets that can be used for experimentation purposes in this domain. We also plan to extend the proposed system using additional baseline metrics and test the impact of exploiting the proposed model on their quality. In addition, we plan to practically test the proposed model with a group of advertisers in order to figure out their feedback on the quality of the suggested candidate advertising terms.

Acknowledgement

The authors are thankful to the Arab American University of Palestine for the fund and support received during the work on this research project.

References

[1] G. Xu, Z. Wu, G. Li, and E. Chen, "Improving contextual advertising matching by using Wikipedia thesaurus knowledge," *Knowledge and Information Systems*, vol. 43, no. 3, pp. 599-631, 2015.Article (CrossRef Link)

[2] H. Choi, C. F. Mela, S. Balseiro, and A. Leary, "Online display advertising markets: A literature review and future directions," *Columbia Business School Research Paper*, no. 18-1, 2019. Article (CrossRef Link)

[3] I. Lopez-Gazpio, M. Maritxalar, M. Lapata, and E. Agirre, "Word n-gram attention models for sentence similarity and inference," *Expert Systems with Applications*, vol. 132, pp. 1-11, 2019.Article (CrossRef Link)

[4] M. Maree and M. Belkhatir, "Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies," *Knowledge-Based Systems*, vol. 73, pp. 199-211, 2015.Article (CrossRef Link)

[5] L. Gao, K. Dai, L. Gao, and T. Jin, "Expert knowledge recommendation systems based on conceptual similarity and space mapping," *Expert Systems with Applications*, vol. 136, pp. 242-251, 2019.Article (CrossRef Link)

[6] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel, "Web page summarization for just-in-time contextual advertising," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 1, p. 14, 2011.Article (CrossRef Link)

[7] Z. Wu, G. Xu, Y. Zhang, P. Dolog, and C. Lu, "An improved contextual advertising matching approach based on wikipedia knowledge," *The Computer Journal*, vol. 55, no. 3, pp. 277-292, 2011.Article (CrossRef Link)

[8] A. Ghose, H. E. Kwon, D. Lee, and W. J. I. S. R. Oh, "Seizing the commuting moment: Contextual targeting based on mobile transportation apps," *Information Systems Research*, vol. 30, no. 1, pp. 154-174, 2019.Article (CrossRef Link)

[9] J.-Y. Chen, H.-T. Zheng, Y. Jiang, S.-T. Xia, and C.-Z. Zhao, "A probabilistic model for semantic advertising," *Knowledge and Information Systems*, vol. 59, no. 2, pp. 387-412, 2019/05/01 2019.Article (CrossRef Link)

[10] Yuping Liu-Thompkins, "A decade of online advertising research: What we learned and what we need to know," *International Journal of Advertising*, vol. 48, no. 1, pp. 1-13, 2019. Article (CrossRef Link)

[11] A. Kumar, A. Nayyar, S. Upasani, and A. Arora, "Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising," *Data Management, Analytics and Innovation*, pp. 3-13, 2019. Article (CrossRef Link)

[12] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting Wikipedia as external knowledge for document clustering," in *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 389-396, 2009. Article (CrossRef Link)

[13] A. N. Pak and C.-W. Chung, "A wikipedia matching approach to contextual advertising," *World Wide Web*, vol. 13, no. 3, pp. 251-274, 2010.Article (CrossRef Link)

[14] K. Yeun Chun, J. Hee Song, C. R. Hollenbeck, and J.-H. Lee, "Are contextual advertisements effective? The moderating role of complexity in banner advertising," *International Journal of Advertising*, vol. 33, no. 2, pp. 351-371, 2014.Article (CrossRef Link)

[15] G. Armano, A. Giuliani, and E. Vargiu, "Studying the impact of text summarization on contextual advertising," in *Proc. of 2011 22nd International Workshop on Database and Expert Systems Applications*, pp. 172-176, 2011. Article (CrossRef Link)

[16] P. Liu, J. Azimi, and R. Zhang, "Automatic keywords generation for contextual advertising," in *Proc. of the 23rd International Conference on World Wide Web*, pp. 345-346, 2014. Article (CrossRef Link)
[17] K. Zhang and Z. Katona, "Contextual Advertising," *Marketing Science*, vol. 31, no. 6, pp. 980-994, 2012.Article (CrossRef Link)
[18] E. Vargiu, A. Giuliani, and G. Armano, "Improving contextual advertising by adopting collaborative filtering," *ACM Transactions on the Web (TWEB)*, vol. 7, no. 3, p. 13, 2013. Article (CrossRef Link)
[19] Z. Wu, G. Xu, R. Pan, Y. Zhang, Z. Hu, and J. Lu, "Leveraging Wikipedia concept and category information to enhance contextual advertising," in *Proc. of the 20th ACM international conference on Information and knowledge management*, pp. 2105-2108, 2011. Article (CrossRef Link)
[20] K. S. Dave and V. Varma, "Pattern based keyword extraction for contextual advertising," in *Proc. of the 19th ACM international conference on Information and knowledge management*, pp. 1885-1888, 2010. Article (CrossRef Link)
[21] J.-H. Lee, J. Ha, J.-Y. Jung, and S. Lee, "Semantic contextual advertising based on the open directory project," *ACM Transactions on the Web (TWEB)*, vol. 7, no. 4, p. 24, 2013. Article (CrossRef Link)
[22] A. Pak, "Using wikipedia to improve precision of contextual advertising," in *Proc. of Language and Technology Conference*, pp. 533-543, 2009. Article (CrossRef Link)
[23] W.-J. Ryu, J.-H. Lee, and S. Lee, "Utilizing verbal intent in semantic contextual advertising," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 7-13, 2017.Article (CrossRef Link)
[24] M. Grbovic *et al.*, "Scalable semantic matching of queries to ads in sponsored search advertising," in *Proc. of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 375-384, 2016. Article (CrossRef Link)
[25] A. Patil, K. Dave, and V. Varma, "Leveraging latent concepts for retrieving relevant ads for short text," in *Proc. of European Conference on Information Retrieval*, pp. 780-783, 2013. Article (CrossRef Link)
[26] J. Zhang and D. Qiao, "A Novel Keyword Suggestion Method for Search Engine Advertising," *PACIS 2018 Proceedings*, 305, 2018.Article (CrossRef Link)
[27] T. Dong and P. Shao, "The Semantic Analysis of Ambiguity in Advertisements," in *Proc. of 2016 5th International Conference on Social Science, Education and Humanities Research*, 2016. Article (CrossRef Link)
[28] H.-T. Zheng, J.-Y. Chen, and Y. J. I. S. Jiang, "An ontology-based approach to Chinese semantic advertising," *Information Sciences*, vol. 216, pp. 138-154, 2012. Article (CrossRef Link)
[29] F. Mahdisoltani, J. Biega, and F. M. Suchanek, "YAGO3: A Knowledge Base from Multilingual Wikipedias," in *Proc. of International Semantic Web Conference*, pp. 177-185, 2015. Article (CrossRef Link)
[30] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?," *Communications of the ACM*, vol. 35, no. 12, pp. 29-38, 1992. Article (CrossRef Link)
[31] G. A. Miller, "WordNet: a lexical database for English," in *Proc. of HLT '93: Proc. of the workshop on Human Language Technology*, p. 409, 1993. Article (CrossRef Link)
[32] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries," in *Proc. of the international conference on intelligent analysis*, vol. 2, no. 6, pp. 2-6, 2005.
[33] M. P. Sinka, D. W. J. S. c. s. d. Corne, management, and applications, "A large benchmark dataset for web document clustering," vol. 87, pp. 881-890, 2002.
[34] G. Armano, A. Giuliani, and E. Vargiu, "Novel Text Summarization Techniques for Contextual Advertising," in *Proc. of the 2nd Italian Information Retrieval (IIR) Workshop, Milan, Italy*, 2011. Article (CrossRef Link)
[35] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13-18, 2013.Article (CrossRef Link)

[36] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel, "Just-in-time contextual advertising," in *Proc. of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 331-340, 2007. Article (CrossRef Link)
[37] G. Armano, A. Giuliani, A. Messina, M. Montagnuolo, and E. Vargiu, "Experimenting Text Summarization on Multimodal Aggregation," in *Proc. of 4th International Work-shop DART 2011, New Challenges on Information Retrieval and Filtering*, 2011.

Mohammed Maree received the Ph.D. degree in information technology from Monash University. He has published articles in various high-impact journals and conferences, such as ICTAI, Knowledge-Based Systems, and the Journal of Information Science. He is currently a Committee Member/Reviewer of several conferences and journals. He has supervised a number of master's students in the field of knowledge engineering, data analysis, information retrieval, natural language processing, and hybrid intelligent systems. He began his career as a Research and Development Manager with gSoft Technology Solution Inc. Then, he worked as the Director of Research and QA with Dimensions Consulting Company. Subsequently, he joined the Faculty of Engineering and Information Technology (EIT), Arab American University, Palestine (AAUP), as a full-time Lecturer. From September 2014 to August 2016, he was the Head of the Multimedia Technology Department, and from September 2016 to August 2018, he was the Head of the Information Technology Department. In addition to his work at AAUP, he worked as a Consultant for SocialDice and Dimensions Consulting companies. He is also the Head of the Multimedia Technology Department, Faculty of Engineering and Information Technology, AAUP.

Rami Hodrob is an assistant professor at Information Technology Department of the Arab American University Palestine (AAUP) and currently the Assistant of VP academic affairs. He was the Head of Computer Technology Information Department in the Faculty of Engineering & Information Technology in AAUP from September 2014 till September 2015 and from September 2018 till September 2019. He is currently working on Erasmus + project in the field of CBHE (TESLA). He received his Ph.D. in 2016 from the Czech University of Life Sciences-Prague and his work focused on Information and communication technology economics. He received the M.Sc. degree in computing where he worked on using graphically notations in ontology engineering from Birzeit University, Palestine, Master of Business Administration degree from An-Najah University, Palestine and B.Sc. in Electronic Engineering from Yarmouk University Jordan in 2012 and 2003, 1995 respectively. He worked as a lecturer at Information Technology Department - Faculty of Engineering and Information Technology - at the Arab American University, Palestine since 2003. Dr. Hodrab has authored many publications in international journals and refereed conference proceedings (http://www.aaup.edu/rami.hodrob)(https://scholar.google.de/citations?user=YFMIRCYAAAAJ&hl=en). He is also a committee member/reviewer of several conferences and journals such as IEEE International Conference on Teaching, Assessment and Learning for Engineering 2012, ALE 2013, IEEE EDUCON 2018 and American Journal of Applied Sciences, 2016. His main research interests are in the fields of ICT Economics, Knowledge Engineering and Ontologies, Semantic Web and Augmented reality.

Mohammed Belkhatir graduated from the University of Grenoble, France with an M.Phil and a Ph.D in Computer Science, both of which were supported by research grants from the French Ministry of Research. He is now an Associate Professor at the University of Lyon, France.

Saadat M. Alhashmi received his PhD from Sheffield Hallam University, Sheffield, UK. Over the years, he has supervised a number of PhD students and published extensively in various high impact journals and conferences. Saadat joined the University of Sharjah, Sharjah, UAE in 2015 as an Associate Professor of MIS. His current research interests are business analytics, big data and the impact of technology on businesses.