Assessing Ontologies Usage Likelihood via Search Trends[1]

Mattia FUMAGALLI [a2], Tania BAILONI [b] and Fausto GIUNCHIGLIA [b]
[a] *Conceptual and Cognitive Modeling Research Group (CORE), Free University of Bozen-Bolzano, Bolzano, Italy*
[b] *Department of Information Engineering and Computer Science (DISI) - University of Trento, Italy*

Abstract. The generation of *high quality* and *re-usable ontologies* depends on effective methodologies aimed at supporting the crucial process of identifying the ontology requirements, in terms of the *number of potential end-users* and *end-users' informational needs*. It is widely recognized that the exploitation of *competency questions* (CQs) plays an important role in this requirement definition phase. In this paper, we aim at introducing a new general approach to exploit *(web) search trends*, and the huge amount of searches that people make every-day with web search engines, as a pivotal complementary source of information for the identification of informal needs of large numbers of end-users. To achieve this goal we use the "auto-suggest" results provided by search engines like *Bing* and *Google* as a goldmine of data and insights. We select a set of keywords to identify the ontology terminology, and we collect and analyze a huge amount of *web search queries* (WSQs) related to the selected set of keywords. In turn, we identify the search trends related to the collected WSQs and we show how the corpus of selected WSQs can be used to assess the *usage likelihood* of a selected ontology w.r.t. the identified (web) search trends. The experimental results are used to discuss the practical utility of the proposed approach.

Keywords. ontologies, web search, search trends, ontologies design, topic modeling, usage likelihood

1. Introduction

Ontologies are the main backbone structure of many semantic applications and are central in supporting semantic interoperability. Building useful, high-quality and re-usable ontologies is not a trivial task and mainly depends on effective methodologies aimed at supporting the process of ontology engineering [1]. In this respect, within the ontology requirements definition phase, which is essential for the ontology development lifecycle [2], the identification of *the amount of possible end-users* of the ontology and *the end-users' informational needs*, is one of the most pivotal activities. Understanding *how many* users may need the ontology and the *users' view* of the knowledge to be encoded in the ontology is key for defining the function of the ontology and enabling its re-usability.

[2] This paper was written under the contract with the University of Trento

In the work of devising methodologies for ontology development, the key role of *competency questions* (CQs) for identifying ontology requirements is widely recognized [3,4]. CQs are specific questions about a given domain of information, and represent valuable information sources for checking the scope of the ontologies being developed.

In a similar spirit, the work presented in this paper aims at devising a methodology to exploit the huge amount of searches that people make every-day with large-scale cross-domain web search engines as a valuable source of information for the ontologies requirements definition phase.

Nowadays people use *web search engines* to find information about almost everything. An outstanding example is provided by Google[3], which processes over 3.5 billion searches per day and 1.2 trillion searches per year worldwide[4]. These searches provide insights about multiple and different people's needs and can be analyzed and exploited in different ways. The words or strings of words that web search engine users type into a search box are called *web search queries* (WSQs). These queries are very different from the ones provided in standard query languages. They are indeed often given in *plain text* with optional "search-directives" (such as "and"/"or" with "-" to exclude), but they are not constrained by fixed syntax rules, as command languages with special parameters [5]. WSQs can be categorized into three main broad categories [6], namely, *informational*, *navigational*, and *transactional* (or "do", "know", "go" [7]). This classification was empirically derived through the analysis of the queries of some of the most used search engines [8] and shows how WSQs are real-world applications of different kinds of natural language questions, providing hints about different motivations and semantic needs.

WSQs encode people *search interest trends*. They represent a goldmine of insights for today's knowledge engineers and can complement the role of CQs in identifying users' view and their semantic needs. Moreover, WSQs can provide, statistically relevant information about the ontology *usage likelihood*, namely the amount of possible end-users of the ontology. This aspect is central to support the re-usability of the ontology. Finally, WSQs can be easily collected by analyzing the suggested results of the most used search engines.

Following the lack of work in exploiting WSQs for ontology (*"data-driven"*) assessment and development [9], we propose a new approach to support the ontology requirements specification phase. We select a set of keywords (KWDs) to identify the ontology terminology. We collect and analyze a huge set of WSQs related to the selected keywords and, in turn, we identify the search trends related to the collected WSQs. As final step, we show how the corpus of selected WSQs can be used to assess a given ontology w.r.t. the identified *(web) search trends*.

The main contributions are:

- given a set of selected keywords, a *procedure* for *gathering, processing* and *analyzing* the WSQs provided by the auto-suggested results of search engines like Google and Bing[5] (Section 2 and 3);
- given a set of selected WSQs, *a machine learning based pipeline* for *identifying web search trends* (Section 4);

[3] `https://www.google.com/`

[4] `https://www.internetlivestats.com/google-search-statistics/`

[5] `https://www.bing.com/`

- (the beginning of) a *search driven ontology assessment method* where we provide a first test, by assessing the usage likelihood of 8 state-of-the-art (SoA) ontologies against a gold-standard data set, manually created by us, with around 8,000 WSQs, grouped by 36 input core keywords and associated to the 8 input SoA ontologies (Section 5).

The paper is completed as follows: Section 6 describes the related work and Section 7 discusses conclusions and future work.

2. Ontologies and Search Trends

Suppose that there is a need to capture the requirements, in terms of semantic needs and user view, of the knowledge to be encoded in an ontology to search for information in the *travel domain*. This activity requires that knowledge engineers take into account somehow the end-users needs, and it is necessary to ground the ontology building process [2,4]. A first central task consists in identifying the objectives and the purpose of the ontology from a user's point of view, namely to determine the domain of interest and the scope. For instance, the ontology may be used to find the best *flights* in terms of *price*, or it may be used to find the *less crowded places* where to go in a certain *period*, and so forth. This identification step involves the selection of a *lexicon* for the given ontology, namely a set of core keywords that will be central in the final application and will be used to identify the questions that users may run over the system. In the context of the travel domain, for instance, we may have keywords like *booking*, *place*, *vacation*, *route*, *seat number*, *period*, *destination*, *hotel*, *price* and *guide*.

The keywords for the determination of the ontology purpose are usually collected in two ways: *i.* from application-specific documents, usually by running knowledge extraction tools [10]; *ii.* directly from the suggestions of the possible end-users, usually through interviews. Another source of information for selecting the lexicon of an ontology, especially if this ontology will be implemented to support mainstream applications (e.g., ecommerce websites) can be the *web search data* offered by free online services like *Google Trends*[6]. These services can provide, given a domain of interest, multiple suggestions and related keywords relying on the huge number of searches that people run every day over mainstream search engines. Selecting the concepts of the ontology by using keywords suggestion tools like *Google Trends* helps in intercepting people's informational needs about a given *selected topic*. This way of deriving the ontology lexicon, while being complementary to the more traditional ones (see *i.* and *ii.* above), introduces two main advantages as well. Firstly, extracting keywords from search trends allows collecting new useful insights about the *number of searches for a given term* (for instance it may be possible to give more weight to the keyword "travel", instead of "vacation", given the much higher number of searches about travels). Secondly, search trends suggestion tools provide a *highly scalable means* by which information about the domain of interest and scope of the ontology can be collected (notice that web searches are not made by experts, and these data should not be considered as a replacement of data gathered from domain experts).

Another central task in the identification of the requirements, which directly follows the keyword selection task, consists of identifying the questions that users may run over

[6] `https://trends.google.com/trends`

the ontology. Similarly to the selection of the lexicon, the suggestion tools provided by mainstream search engines can play a pivotal role. Along with the competency questions (CQs) [3] provided by domain experts and possible end-users, insights about WSQs, related to the selected keywords, can be used as a complementary source of information to specify the semantic needs and user views of the knowledge to be encoded in the ontology, and to assess the *usage likelihood of an ontology*.

Table 1. Examples of WSQs for the "travel" domain

WSQ1	are flights still going to china
WSQ2	are flights cheaper on boxing day
WSQ3	when flights get delayed
WSQ4	can flights be cancelled due to snow
WSQ5	flights where you can choose seats
WSQ6	which flights allow pets
WSQ7	why flights to brazil are so expensive
WSQ8	what flights go from terminal 2 manchester

WSQs are essentially questions at a conceptual level an ontology may be able to answer. For instance, the ontology may be used to address questions like *"are flights still going to China?"*, or *"can flights be canceled due to snow?"* (see Table 1 for more examples). Differently from competency questions, which are identified through interviews with domain experts and end-users brainstorming, WSQs are derived by analyzing the suggestion results of *general* (e.g., Google and Bing) or *vertical* search engines (e.g., *Library of Congress* or *Nuora*[7]).

WSQs represent a significant corpus of information from which users' view and usage likelihood of the knowledge to be encoded in the ontology can be extracted. In this paper, we categorize this corpus of information in *(web) search trends*, namely set of *weighted keywords* derived from *sets of web search queries*. WSQs can be then used during the test workflow to assess the ontology w.r.t some given *(web) search trends*. Using WSQs to identify search trends and assessing ontologies w.r.t. search trends is an open research problem [11]. WSQs, besides being written in natural language, i.e., plain text, are indeed very noisy and not constrained by fixed syntax rules. We model this problem as a process where the main goal is to calculate the *similarity* of an ontology vocabulary w.r.t. a selected corpus of (web) search trends. The focus of our approach is mainly to support ontology engineers in properly verifying: *a.* whether an ontology can be used to address the semantic needs expressed by the identified search trends; *b.* the usage likelihood of the ontology.

3. From Keywords Selection to WSQs

Let us imagine, that a knowledge engineer has to develop a *semantic web application* (i.e., a web service like an ecommerce website, a vertical search engine on a website, and so forth) and needs to determine the knowledge and information that are necessary for structuring the semantic of the data with an ontology. The first task she has to address is to identify a corpus of keywords and queries that concerns the target domain. Once the

[7] `https://www.nuroa.co.uk/`, `https://www.loc.gov/`

corpus is determined, she is able to develop from scratch the ontology or to select from a number of existing ontologies the most appropriate for the application.

3.1. Gathering Data

Services that provide data and insights about searching keywords and web search queries are pivotal means by which the initial data selection process could be facilitated. In what follows, through a running example, we describe how we addressed the task of gathering and analyzing WSQs given a set of keywords.

As first step, in order to make the approach as general as possible, we focused on very broad domains of interest, and we investigated what can be the most searched and commonly used keywords over the web. As input references we considered:

- the most commonly used entities in the *Google Knowledge Graph*[8] (namely the knowledge graph used by Google and its services to enhance search engine's results);
- Rosch's [12] work on *basic level categories*, which provides a cognitive grounding for the selection of the most informative categories and related keywords;
- the *Schema.org*[9] vocabulary, and core entities, itself related to the Google knowledge graph, being one of the most used semantic resources for structuring and semantically enriching web documents content.

After the analysis of these resources, we collected 36 keywords. 29 of them were identified within the *common Schema.org types* used by the *Knowledge Graph Search API*, namely: *Action; Book Series; Book; Creative Work; Educational Organization; Event; Government Organization; Local Business; Movie Series; Movie; Music Album; Music Group; Music Recording; Offer; Organization; Periodical; Person; Place; Product; Recipe; Restaurant; Review; Sports Team; TV Episode; TV Series; Vehicle; Video Game Series; Video Game; Website*. 7 of them were identified within Rosch's basic level categories, namely: *Bird; Clothing; Fish; Fruit; Furniture; Tool; Tree*. Notice that we produced the final list with the purpose of covering a very broad set of common sense everyday searches, without claiming to be exhaustive.

As second step, starting from the selected keywords, we collected the corresponding WSQs. In order to gather this data we used two main tools, namely *Google Trends* and *Answer The Public*[10]. Google Trends provides reliable and updated insights about Google searches, while *Answer The Public* offers a very easy-to-use (free) service to collect a good amount of information about web searches, by using Google and Bing APIs. Both of these services, given a set of keywords, provide the corresponding *"most typed" WSQs*. Moreover, *Answer The Public* organizes WSQs according to different criteria, or "modifiers". For instance, WSQs can be categorized per *keyword*, *typology*, e.g., 'questions' or 'comparison', or *types of adverb*, like 'what' or 'where'. This organization can be used as a starting point for a fine-grained categorization of WSQs (for instance, WSQs can be grouped as 'temporal' or 'spatial', according to their corresponding modifiers). By merging and processing (e.g., we needed to delete some noisy or irrelevant WSQs) the results gathered with the above-introduced services, we collected around 8,000 WSQs.

[8] `https://developers.google.com/knowledge-graph`
[9] `https://schema.org/`
[10] `https://answerthepublic.com/`

All these queries have been categorized in relation to the input 36 keywords listed above. Moreover, four main preliminary categories of queries, all of them grouping specific 'modifiers', have been identified:

- *Questions*, i.e., WSQs characterized by modifiers like 'how', 'are', 'what', 'where', 'who', 'which', 'will', 'when', 'can' and 'why';
- *Prepositions*, i.e., WSQs characterized by modifiers like 'for', 'near', 'to', 'with', 'without', 'is' and 'can';
- *Comparisons*, i.e., WSQs used to compare something with something else (e.g., iPhone vs. Samsung), characterized by modifiers like 'and', 'like', 'or', 'versus' and 'vs';
- *Related* and *Alphabeticals*, i.e., general WSQs mainly related to the reference keyword, which cannot properly be identified as from the categories above.

3.2. WSQs Data Set Generation

After grouping all the WSQs for all the selected keywords, we processed the data to decrease the noise. Notice that, at the current state, we performed this step manually, but we are evaluating solutions for automatic support (e.g., by means of *ad hoc* NLP techniques) as part of the future work.

As first step, we checked every single WSQ independently of its categorization. We corrected queries with typos, we dropped duplicated WSQs and WSQs written in languages different from English (the language we selected as reference). For instance, we dropped queries like 'periodical ka hindi' or 'website kaise banaye in hindi' that ask for information in Hindi.

Table 2. Example of WSQs outputs and categorization for the keyword "website"

Keyword	Category	Modifier	WSQ
website	question	are	>*are website expenses deductible*
			>*are website terms of use required*
		can	>*can website access camera*
			>*can website detect vpn*
		when	>*when website was created*
			>*when website was published*
		which	>*which website to watch anime*
		who	>*who website belongs to*
			>*who website is registered to*
	preposition	for	>*website for photographers*
			>*website for selling items*
		is	>*website is under maintenance*
			>*website is not secure*
	comparison	with	>*website with free images*
			>*website with games*
		like	>*website like youtube*
			>*website like airbnb*
	general		>*website visitor counter*
			>*website url*
			>*neargroup website*

As second step, we assessed the WSQs main categories and modifiers. Each single WSQ was checked according to its categorization, namely, we checked whether it was

correctly categorized. Moreover, some groups of WSQs associated with some specific modifiers were discharged because of their incompatibility with the ontology query answering capability. For instance, we dropped those WSQs that imply deep and complex processing like those with modifiers like 'how' and 'why' (e.g., 'how bird eat', 'why fruit is good for you') and those that are pointless like, 'near me', 'furniture is' and 'person is dead'.

The main output is represented by Table 2 (above). Here we have a categorization of some WSQs for the keyword 'website', with related categories and modifiers. The table follows the structure of the categorization: from the general keyword to the specific WSQ.

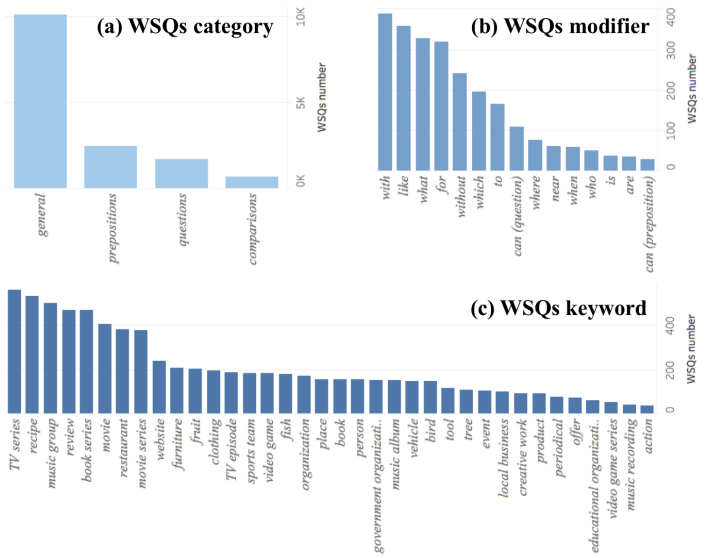

Figure 1. WSQs dataset overview - number of WSQs per (a) category, (b) modifier, and (c) keyword

The distribution of the WSQs, over the whole dataset, given their keywords, categories and modifiers, is shown by Figure 1. From the chart, it is possible to identify the size of the selected keywords, categories and modifiers in terms of the number of associated WSQs. Notice that the number of queries in the different categorizations is quite unbalanced, this being motivated by the available extracted data (e.g., there are more examples of 'general' queries and there are more examples of queries about 'TV-series' than queries about 'trees'). Looking at Figure 1(a) the 'general' category is the broadest; with the 67% of the total number of queries. Figure 1(b) shows the distribution of the queries in terms of modifiers. Modifiers providing comparisons are the most represented across the selected WSQs: 'with' and 'like', for instance, are the broadest groups with 390 and 359 queries respectively, while modifiers like 'are' and 'can' (categorized as prepositions) have only 36 and 31 queries each. Figure 1(c) shows the distribution of the queries among the keyword classification. Most of the queries are related to 'TV series' (562), while 'action', for instance, (clearly a much more abstract keyword) has only 39 related queries.

4. Search Trends Determination

The goal here is *to extract web search trends from the selected WSQs dataset*, where we assume that each set of WSQs associated to a selected keyword consists of a mixture of *web search trends* (WSTs), and a WST[11] is a set of *weighted keywords*. Loosely speaking, our basic assumption is that the semantics of search trends is somehow 'hidden' inside the multiple collected WSQs. As a result, the process of extracting web search trends consists in uncovering this hidden semantics, i.e., trends, that characterize a given set of WSQs. In order to achieve this task we adopted one of the foundational techniques in topic modeling, namely the *Latent Dirichlet Allocation (LDA)* [13] approach. We applied LDA by using the *Parallel Topic Model* of the *Mallet library*[12] [14] with *sparse LDA sampling scheme and data structure* [15]. LDA is a generative probabilistic model that uses *Dirichlet priors* for the *document-topic* and *word-topic* distributions [16]. We considered this widely applied algorithm as a suitable algorithm for achieving our WSTs determination goal (in the context of this current assessment method set-up). Notice that we are aware of other topic-extraction algorithms and of recent developments of LDA (e.g., LDA + embeddings like *lda2vec* [17]), however the evaluation of other topic-extraction algorithms is out of the scope of this paper.

Adapting the formal description of LDA, we modeled a set of documents W as $W = (w_1, ..., w_n)$, each document being a set of *WSQs*. Similarly, we modeled K as the set of all the keywords collected by all the documents W and $K_w = (k_1, ..., k_n)$ as the set of keywords in a document w. Then, given a set of trends $T = (t_1, ..., t_n)$, where a trend is a distribution of keywords in K (e.g., $Creative - Work = \langle 0.3\ Movies, 0.4\ Books, 0\ Fruit, 0.2\ Document, 0.1\ Price \rangle$), we looked for:

(i) *the probability of each trend t_i occurring in document w_i* (from 0 to 1);
(ii) *the weight of each keyword k_i, for a given trend t_i (from 0 to n).*

Starting from the data set of WSQs described in Section 3, in order to identify a group of search trends we adopted the two following approaches:

(a) we manually grouped the WSQs associated to the source keywords into 5 broader documents, associating each of these documents to a trend and then calculating the weight of each keyword for the given trends (from now on we call this approach *"semi-automatic approach"*);
(b) we automatically identified 5 search trends from the complete list of WSQs documents (one per keyword) and then we calculated the weight of each keyword for the given trends (from now on we call this approach *"(fully-) automatic approach"*);

In both cases (*a.* and *b.*) we achieved the goal by addressing the following steps: 1. we took a set of WSQs (associated to a keyword or an arbitrary group) as an input *document*; 2. we processed each document via a NLP pipeline that performs various steps, including: 2(a). tokenization; 2(b). lower case all characters; 2(c). filter out *stop-words*, 2(d). find corresponding *synonyms* and *hypernyms* in WordNet[13] for each extracted term; 3. we applied the LDA algorithm over the whole set of documents to extract the specific

[11] From now on we use 'WST', 'trend' and 'search trend' interchangeably
[12] `http://mallet.cs.umass.edu/diagnostics.php`
[13] `https://wordnet.princeton.edu/`

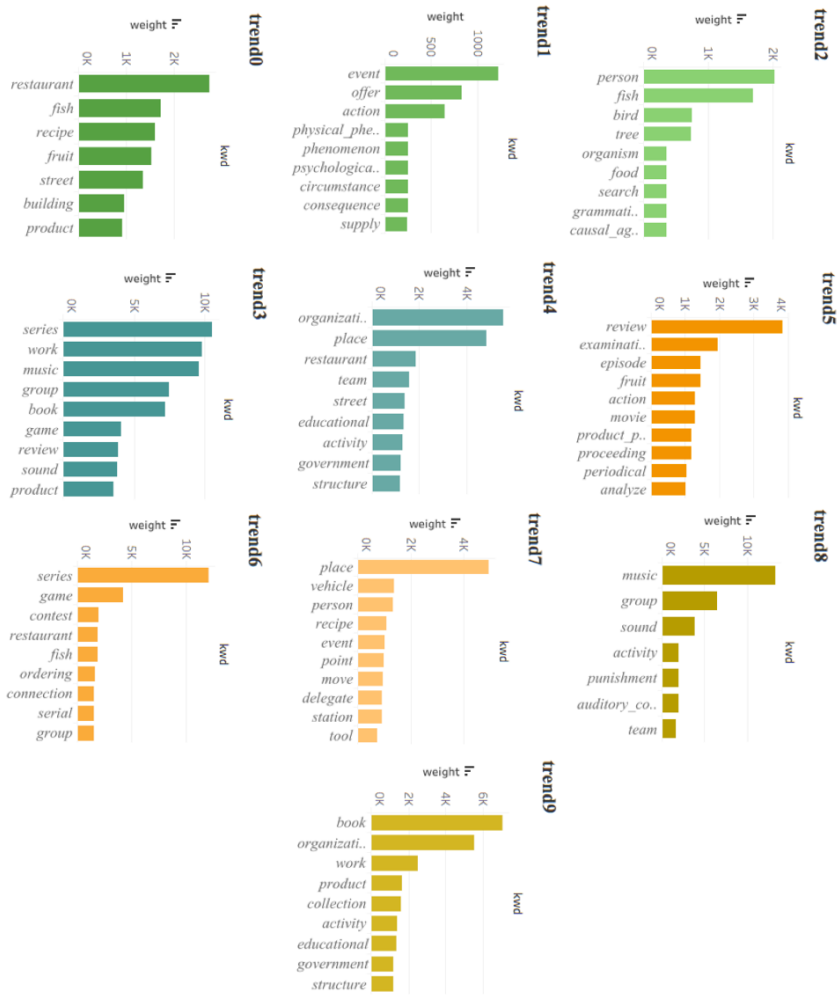

Figure 2. Search trends visualization according to the keyword (kwd) weights generated by means of LDA

probability distribution for 5 topics and the weights of all the keywords (when the input documents are divided per keywords the trends extraction is automatically derived, when the input documents are the 5 groups we manually defined, the trends extraction is mapped to the manual grouping).

The semi-automatic grouping of the documents containing the list of WSQs led to the identification of the following trends:

- **trend$_0$**: *fruit*, *fish*, *restaurant*;
- **trend$_1$**: *offer*, *event*, *action*;
- **trend$_2$**: *person*, *fish*, *bird*, *tree*;
- **trend$_3$**: *creative-work*, *review*, *periodical*, *book*, *book-series*, *music-recording*, *movie-series*, *movie*, *tv-series*, *tv-episode*, *videogame-series*, *videogame*, *clothing*, *furniture*, *website*, *recipe*, *music-album*, *tool*, *vehicle*, *product*, *review*;
- **trend$_4$**: *local-business*, *restaurant*, *place*, *music-group*, *government-organization*, *organization*, *sports-team*, *educational-organization*.

The automatic extraction of 5 trends from the documents containing the list of WSQs, grouped according to the reference keyword, generated the following correlations:

- **trend$_5$**: *product*, *review*, *action*, *movie*, *tv-episode*, *fruit*;
- **trend$_6$**: *bird*, *video-game*, *local-business*, *offer*, *clothing*, *videogame-series*, *movie-series*, *fish*, *restaurant*, *tv-series*, *furniture*;
- **trend$_7$**: *person*, *event*, *place*, *tool*, *recipe*, *vehicle*;
- **trend$_8$**: *music-group*, *music-recording*, *music-album*, *sports-team*, *tree*;
- **trend$_9$**: *website*, *government-organization*, *organization*, *book*, *creative-work*, *educational-organization*, *periodical*, *book-series*.

Each identified trend along with the top keywords is shown in Figure 2. Each semi-automatically derived trend can be described as follows: $trend_0$ is clearly about *food products and facilities*; $trend_1$ is a more abstract trend, highly characterized by *events*, *happenings*, *activities* and other *occurrences*; $trend_2$ groups the main categories of *living beings*; $trend_3$ is clearly about *media* and *creative works*; finally, $trend_4$ is heavily characterized by keywords about *organizations*. In turn, each automatically derived trend can be described as follows: $trend_5$ is very related to *media* and *creative works* with a high focus on the keyword *review*; $trend_6$ is again very related to *creative works* with a strong focus on the keyword *series*; $trend_7$ is heavily related to the keyword *place*; $trend_8$ is focused on *music* and *music groups* and $trend_8$ is clearly related to *organization* and concrete media objects like *books*.

The first observation is that for the semi-automatically generated trends, as expected, we had very coherent groupings, which are influenced by the manual selection of the WSQs before the trend definition phase. The second observation is that the distribution of the weights over the keywords was more balanced for the automatically generated trends. This is because in the semi-automatically generated trends we generated documents with huge amount of WSQs, with multiple overlapping keywords, thus allowing the generation of a broader set of identifying keywords, but, at the same time limiting the "inclusiveness" of the trend (i.e., to be identified as related to one of these trends the weight for the related keyword must be higher).

5. Assessment via Search Trends

After producing a data set of WSQs, given a set of selected input keywords, and after determining a set of WSTs, we are ready to select the ontologies to be assessed. At this point, two are the phases we performed.

Firstly (Section 5.1), we selected a set of SoA ontologies, and: *i.* we generated a corpus, i.e., what we call here "core" data set, from each ontology via a NLP pipeline (as from Section 4); *ii.* we mapped the WSQs of the data set described in Section 3 to the selected ontologies, in order to generate a reference gold standard data set. Notice that the creation of the gold standard is not a mandatory step for the final assessment. However, for the sake of research, we generated this data set in order to better understand the efficiency of the approach we are proposing. Secondly (Section 5.2), we run the assessment and compared the corpus derived from the ontologies concepts with the trends generated through the processes (semi-automatic and automatic) described in Section 4.

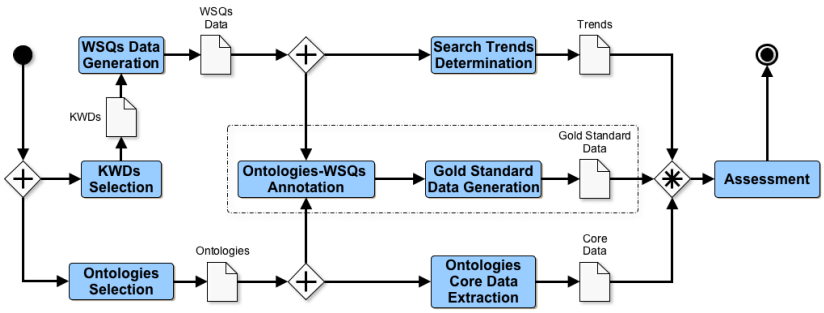

Figure 3. An overview of the process for assessing ontologies via search trends

The results were then tested against the results provided by the gold standard data set derived by means of the previous phase.

Figure 3 provides an overall view of the entire pipeline and shows how the above mentioned phases are combined with the phases described in the previous sections. The blue boxes in the diagram represent the activities we run along the pipeline, the *doc* icons represent the related output. Activities and output grouped by the dotted line represent the phase dedicated to the gold standard generation (to be considered as optional). Notice that, as one of the contributions of this paper, the pipeline we implemented using *RapidMiner*[14] framework and the annotated data set were published free for research purposes.[15]

5.1. Ontologies Core Data Set and Gold-standard Data Set Generation

As first activity, we assessed most of the available ontologies by looking at the *Linked Open Vocabulary Catalog* (LOV)[16] and other sources (see for instance *Datahub*[17]). Following the running example introduced by the previous sections, we took into account both *general-purpose* and *domain-specific* ontologies, where the former are better in answering queries from a wide range of subjects, while the latter are better in answering queries of particular and domain-specific subjects. For the selection of the domain ontologies we considered the most specific keywords from the list defined in Section 3 (e.g., 'movie', 'tv episode', 'tv series' or 'local business' and 'offer'). At the end of the analysis process we identified four generic (or general-purpose) and four domain-specific ontologies, namely: *Schema.org*, *Opencyc*[18], *SUMO*[19], *DBpedia*[20], *GR*[21], *EBUCore*[22], *BioTop*[23], *MO*[24].

[14] http://www.rapidminer.com
[15] https://github.com/Matt-81/ontologies-by-trends
[16] https://lov.linkeddata.es/dataset/lov
[17] https://datahub.io/
[18] https://old.datahub.io/dataset/opencyc
[19] http://www.adampease.org/OP/
[20] http://dbpedia.org/ontology/
[21] http://www.heppnetz.de/ontologies/goodrelations/v1
[22] https://www.ebu.ch/metadata/ontologies/ebucore/index.html
[23] http://biotopontology.github.io/
[24] http://musicontology.com/

Table 3. Distribution of WSQs over ontologies

(a)

Ontology	Matches no.
Schema.org	*5332*
DBpedia	*2919*
OpenCyc	*1095*
SUMO	*1169*
BioTop	*96*
GR	*172*
EBUCore	*283*
MO	*491*

(b)

Ontologies no.	WSQs no.
0	*1097*
1	*3247*
2	*1323*
3	*1503*
4	*298*
5	*86*
6	*7*
7	*0*

After the identification of the candidate ontologies, we addressed the mapping with the given input WSQs to generate the gold standard data set[25]. We analyzed every single ontology by using Protégé[26], this was in order to check whether each ontology contains concepts or properties that can be mapped to the given input keywords coming from every single WSQ. The modifiers associated with the WSQs were particularly useful to check for objects or data properties. For instance, the WSQ 'product like kindle', asking for products that are similar to the *Kindle product*, can be mapped into *schema.org* with the property 'isSimilarTo', which has both *domain* and *range* in the classes 'Product' and 'Service'. This WSQ can be also mapped into *DBpedia*, *OpenCyc* and *GR* for similar reasons. To give another example, we mapped the query 'who organization members', which asks for the members of the WHO *(World Health Organization)*, to *schema.org*, because of the property 'member', which has *domain* in the class 'Organization' and *range* in 'Organization' and 'Person' classes. This mapping is feasible also for *DBpedia*, thanks to the property 'organisation member' that has *domain* in the class 'Organisation' and *range* in 'Organisation member'. The same can be said for the other ontologies, i.e., *OpenCyc*, *SUMO*, *EBUCore*, where the mapping is supported by 'member' properties. At the end of the mapping process, we generated a data set with very fine-grained information about which ontology (within the analyzed ones) can be used to answer the collected WSQs. The main insight is that the 85% of the collected WSQs have a match in at least one of the selected ontologies.

Table 3 shows the number of matches found for each ontology *(a)*, and the number of queries grouped by the number of different ontologies mapped *(b)*. Firstly, as expected, it can be noticed that, considering the distribution of the matches across the ontologies, the general-purpose ontologies (i.e. *schema.org*, *DBpedia*, *OpenCyc* and *SUMO*) have the highest coverage and *schema.org* is the topmost in the ranking. Moreover, it can be observed that 1097 suggestions (15%) cannot be matched. The majority of the queries (3247 WSQs, 43%) have only one match, however many of them can be mapped over two or three ontologies (respectively the 20% and 17%). It can be further noticed that the maximum number of ontologies with a "query match" is six, thus no suggestion can be found in all the eight ontologies studied. Similarly, the analysis of the mapping of the WSQs grouped by the keyword categorization highlighted the difference in the distribution of the matches. For instance, all the WSQs with the keyword 'music recording' found a match, while less than half of those with the keyword 'clothing' can be an-

[25] For the gold standard generation task, a master student and postdoctoral researcher from computer science, with high expertise in knowledge engineering were involved

[26] `https://protege.stanford.edu/`

swered using the ontologies selected. Analyzing the distribution of the WSQs, grouped by keyword, over the ontologies was helpful to understand the "coverage" of each keyword among the ontologies (i.e., which ontology maps a specific keyword), in fact, some keywords have matches in only one ontology (i.e. 'clothing, 'furniture', 'local business', 'movie series' and 'recipe'); while others are present in more than one (e.g. keyword 'fruit' can be found in *OpenCyc* and *SUMO*).

As final step, we extracted the core data set and the gold standard data set, where: *i.* the former is generated from a set of text files, each one representing an ontology and collecting the related *triples* with class-properties relations (e.g., 'Person'- 'DomainOf'- 'hasName'; 'Album'- 'DomainOf'-'hasAuthor'); *ii.* the latter, is generated from a set of text files, each one representing an ontology and collecting the list of WSQs that can be addressed by that ontology according to the above described annotation task.

We then ran the extraction of the two data sets by applying the same NLP pipeline we used to derive the web search trends (see Section 4), and, for each ontology, we generated a corpus of weighted keywords from the corresponding triples and a corpus of weighted keywords from the annotation.

5.2. Assessing ontologies

The goal here is to run a preliminary experiment to *assess ontologies w.r.t. a given set of search trends*, and thus trying to infer their usage likelihood. In order to achieve this goal, we ran two main experimental trials. The first one was to evaluate the ontologies w.r.t. the semi-automatic generated web search trends, i.e., $trend_0$, $trend_1$, $trend_2$, $trend_3$, $trend_4$. The second was to evaluate the ontologies w.r.t. the automatic generated web search trends, i.e., $trend_5$, $trend_6$, $trend_7$, $trend_8$, $trend_9$. In both the trials, the corpus derived directly from the ontologies was tested against the corpus generated from the gold standard.

The results of the first and the second trials are showed in Table 4. In the extreme left column, the input corpus for each ontology is reported. The manually annotated golden standard data sets are denoted as *"-gold"*, the data sets generated directly from the ontology are denoted as *"-core"*. All the other columns report a confidence prediction value (from 0 to 1) for each reference trend. We take this confidence as the measure to assess the given input ontology (i.e., the *probability of a trend occurring in an ontology*, see Section 4).

The first observation concerns the comparison between the "-gold" and "-core" data sets prediction results. In both the trials, the ontologies for which the "-gold" prediction value is aligned to the "-core" value are *DBpedia*, *EBUcore* and *mo*. This means that about the 60% of the "-core" predictions is not aligned with the "-gold" predictions, thus highlighting a relevant difference between the "-gold" and "-core" corpora in general (see for instances the example of *opencyc*, where this difference is very high).

The second observation concerns the differences between the results related to the automatically generated trends and the semi-automatically generated trends. The ontologies of the first trial were associated mostly to $trend_0$ and $trend_1$. The ontologies of the second trial were mostly associated to $trend_6$ and $trend_9$. By looking at the keywords composing the generated trends (Section 4) it is possible to notice that the most selected trends, in both the trials, cover a wide range of keywords, from 'food', 'living being' and 'event', for the semi-automatically generated trends, to 'organization' and 'creative work' for the automatically generated trends. Still, the prediction results of the second

Table 4. Ontology assessment prediction values, given the selected trends

corpus	trend0	trend1	trend2	trend3	trend4	trend5	trend6	trend7	trend8	trend9
biotop-core.txt	0,14	0,70	0,06	0,08	0,02	0,08	0,26	0,11	0,16	0,38
biotop-gold.txt	0,67	0,21	0,00	0,12	0,00	0,06	0,01	0,31	0,60	0,02
dbpedia-core.txt	0,09	0,88	0,03	0,00	0,00	0,09	0,16	0,22	0,12	0,40
dbpedia-gold.txt	0,14	0,45	0,04	0,23	0,15	0,26	0,22	0,10	0,11	0,31
ebucore-core.txt	0,00	1,00	0,00	0,00	0,00	0,00	0,16	0,05	0,03	0,76
ebucore-gold.txt	0,04	0,61	0,14	0,07	0,15	0,13	0,32	0,00	0,06	0,50
gr-core.txt	0,07	0,75	0,00	0,06	0,12	0,02	0,24	0,12	0,16	0,45
gr-gold.txt	0,40	0,17	0,00	0,35	0,08	0,28	0,15	0,02	0,41	0,14
mo-core.txt	0,00	1,00	0,00	0,00	0,00	0,00	0,41	0,10	0,01	0,48
mo-gold.txt	0,07	0,65	0,02	0,26	0,00	0,34	0,20	0,02	0,01	0,44
opencyc-core.txt	0,01	0,27	0,62	0,10	0,00	0,06	0,38	0,26	0,04	0,26
opencyc-gold.txt	0,50	0,16	0,00	0,34	0,01	0,28	0,15	0,08	0,39	0,10
schema-core.txt	0,65	0,35	0,00	0,00	0,00	0,04	0,13	0,06	0,61	0,15
schema-gold.txt	0,17	0,53	0,02	0,13	0,15	0,21	0,19	0,08	0,17	0,34
sumo-core.txt	0,13	0,63	0,15	0,09	0,00	0,16	0,26	0,19	0,11	0,28
sumo-gold.txt	0,39	0,15	0,04	0,30	0,12	0,27	0,12	0,04	0,35	0,23

trial are less unbalanced than the prediction results of the first trials (see values across trends).

The third observation concerns the association between ontologies and trends. As we expected, the general ontologies are usually associated to broad trends, i.e., trends with a wide range of keywords. See for instance the example of *DBpedia* and *Schema.org*, which are associated with trends like $trend_0$, $trend_1$ and $trend_9$. On the other hand, for what concerns the more specific and domain-oriented ontologies, it is possible to notice that there can be a mismatch between their target and the trends they are associated with. One familiar with the *Music Ontology*, for instance, would have expected a strong connection with $trend_3$, $trend_4$ or $trend_8$. However, *mo* is associated with $trend_1$ and $trend_9$ (and that is very strong compared to the expected trends).

While we are careful not to draw overly general conclusions from this preliminary experiment with a small set of ontologies, we still observe a few salient phenomena. First of all, the differences between the "-core" and "-gold" results, suggest a further analysis of how the "-core" corpora can be directly extracted from the ontologies. The challenge here is to properly identify the choices applied during the annotation process and the generation of the gold standard, and to devise the proper automatic solution to better simulate that choices. In the above experiment, for instance, we implemented the NLP pipeline we described in Section 4. More NLP options need to be considered and compared. The similar impact of the semi-automatic grouping and the automatic grouping suggests, on the other hand, that: *i.* grouping manually WSQs does not necessarily imply a more understandable or clear-cut division of trends, rather it may bias the generation process (the confidence values for the semi-automatically generated trends are indeed more unbalanced, compared to the automatically generated trends; see, for instance $trend_1$ vs. $trend_2$); *ii.* the WSQs preprocessing phase, where WSQs may be 'cleaned' or 'deleted', has a central role in the generation of trends, thus multiple versions of WSQs text files, created after the keywords' selection step, should be tested and compared. Finally, a more efficient set-up for the generation of trends might be found, and different approaches should be tested and compared (see for instance the integration of LDA with embeddings [17]). We foresee these improvements of our setup as immediate future work.

6. Related work

Our work is mainly related to the huge research effort in *ontology (functional) evaluation* [18,19,20,21]. More specifically, the work that most overlaps with our efforts is that on *data driven* and *competency questions (CQs) ontology evaluation* [9,3], where the main goal is to support the ontology development requirement specification phase and facilitating the reuse of these semantic data structures. This work has been extensive and has exploited a huge amount of methods and techniques including, e.g. *OntoKeeper* [22] and *TONE* [1] (the former being a *semiotic-driven approach* for assessing biomedical ontologies, the latter being a very high precision evaluation method based on the *concepts semantic richness* notion). Our work differs from this in two major respects. The first is that we ground our data-driven approach on information gathered from the search data coming from large-scale web search engines. The formalization and the experimental set-up of our method are then heavily influenced by the nature of this kind of data source and, in particular, by the necessity of adapting and modeling this data in order to make them exploitable in the context of ontology evaluation. Our goal is to propose a practically useful method to extend the assessment and analysis opportunities that are currently available for knowledge engineers. The second difference, which is actually a consequence of the first, is that, by offering a method for exploiting web search data, we allow knowledge engineers to rely on a huge amount of valuable information that can be integrated to the one gathered by means of more traditional methods. This new highly-scalable approach to intercept people's informational needs about specific domains will help indeed the knowledge engineers to understand better what is the real potential of an ontology for a given domain of application.

As last consideration, it is important to observe how the method we applied to identify *web search trends* has been widely studied in the context of *topic modeling* [23,24]. Our approach introduces a new perspective on the application of these approaches over the web search data scenario.

7. Conclusion and Future Work

In this paper, we have proposed a general approach for the assessment of ontologies according to (web) search trends, namely sets of weighted keywords derived from sets of web search queries. This, in turn, has allowed us to better understand how this assessment method can play a central role in the ontology engineering phase, in particular, by supporting the identification of ontologies usage likelihood.

The future work will concentrate on a more fine-grained implementation of the search trends determination phase and an extension of the experimental set-up for improving the understanding of the prediction results and better simulate the results related to the gold-standard data sets. Another future goal is to generate a broad categorization of web search trends, considering a huge amount of search keywords and related domains of interest, and assess most of the currently available ontologies w.r.t. these identified trends.

Acknowledgements. This work has been supported by the project "DELPhi - DiscovEring Life Patterns" funded by the MIUR *Progetti di Ricerca di Rilevante Interesse Nazionale* (PRIN) 2017 - no. 1062, 31.05.2019

References

[1] Demaidi MN, Gaber MM. TONE: A Method for Terminological Ontology Evaluation. In: Proceedings of the ArabWIC 6th Annual International Conference Research Track; 2019. p. 1–10.

[2] De Nicola A, Missikoff M, Navigli R. A software engineering approach to ontology building. Information systems. 2009;34(2):258–275.

[3] Bezerra C, Freitas F, Santana F. Evaluating ontologies with competency questions. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 3. IEEE; 2013. p. 284–285.

[4] Grüninger M, Fox MS. The role of competency questions in enterprise engineering. In: Benchmarking—Theory and practice. Springer; 1995. p. 22–31.

[5] Mihajlovic V, Hiemstra D, Blok HE, Apers PM. Exploiting query structure and document structure to improve document retrieval effectiveness. Centre for Telematics and Information Technology (CTIT); 2006.

[6] Figueroa A. Exploring effective features for recognizing the user intent behind web queries. Computers in Industry. 2015;68:162–169.

[7] Gibbons K. Do, Know, Go: How to Create Content at Each Stage of the Buying Cycle. Search Engine Watch Retrieved. 2014;24.

[8] Jansen BJ, Booth DL, Spink A. Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management. 2008;44(3):1251–1266.

[9] Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data driven ontology evaluation. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). 2004.

[10] Al-Aswadi FN, Chan HY, Gan KH. Automatic ontology construction from text: a review from shallow to deep learning trend. Artificial Intelligence Review. 2019:1–28.

[11] McDaniel M, Storey VC. Evaluating Domain Ontologies: Clarification, Classification, and Challenges. ACM Computing Surveys (CSUR). 2019;52(4):1–44.

[12] Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive psychology. 1976;8(3):382–439.

[13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993–1022.

[14] Newman D, Asuncion A, Smyth P, Welling M. Distributed algorithms for topic models. Journal of Machine Learning Research. 2009;10(Aug):1801–1828.

[15] Yao L, Mimno D, McCallum A. Efficient methods for topic model inference on streaming document collections. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining; 2009. p. 937–946.

[16] Grinshpan AZ. An inequality for multiple convolutions with respect to Dirichlet probability measure. Advances in Applied Mathematics. 2017;82:102–119.

[17] Moody CE. Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:160502019. 2016.

[18] Hlomani H, Stacey D. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. Semantic Web Journal. 2014;1(5):1–11.

[19] Gangemi A, Catenacci C, Ciaramita M, Lehmann J. Modelling ontology evaluation and validation. In: European Semantic Web Conference. Springer; 2006. p. 140–154.

[20] Giunchiglia F, Fumagalli M. Entity Type Recognition – dealing with the Diversity of Knowledge. In: Seventeenth International Conference on Principles of Knowledge Representation and Reasoning; 2020. .

[21] Giunchiglia F, Fumagalli M. Teleologies: Objects, actions and functions. In: International conference on conceptual modeling. Springer; 2017. p. 520–534.

[22] Manion F, Liang C, Harris M, Wang D, He Y, Tao C, et al. OntoKeeper: Semiotic-driven ontology evaluation tool for biomedical ontologists. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 1614–1617.

[23] Wallach HM. Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning; 2006. p. 977–984.

[24] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019;78(11):15169–15211.