Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A tool for recommending keywords with more live and more attention.

Jorge Chamorro-Padial (jorgechp@correo.ugr.es)
University of Granada
Rosa Rodríguez-Sánchez
University of Granada

Research Article

Keywords: Ontology, Attention, Survival, Bibliometrics, Keywords, Papers

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1662994/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A tool for recommending keywords with more live and more attention

Jorge Chamorro-Padial[12]

Rosa Rodríguez-Sánchez[3]

Abstract

In this paper, we propose a method to help authors to choose alternative Keywords that help their papers to gain visibility. These alternative keywords must have a certain level of popularity in the scientific community and, at the same time, be keywords that have fewer competitors. The competitors would be derived from other papers containing the same keywords. Having fewer competitors would allow the author's paper to have a higher consult frequency. In order to recommend keywords, we must first determine an Attention-Survival score. The attention score is obtained by using the popularity of a keyword. The survival score is derived by the number of manuscripts using the same keyword. With these two scores, we created a new algorithm that finds alternative keywords with a high Attention-Survival score. We used ontologies in order to ensure that alternative keywords proposed by our method are semantically related to the original authors keywords that authors wish to refine. The hierarchical structure in an ontology supports the relationship between the alternative keywords and the input keywords. To test the sensibility of the ontology, we used two sources: WordNet and The Computer Science Ontology. Finally, we launched a survey to have human validation for our algorithm, by using keywords from Web of Science papers and three ontologies: WordNet, CSO and DBpedia. We obtained good results in all our tests.

Keywords: Ontology, Attention, Survival, Bibliometrics, Keywords, Papers

Introduction

Authors use keywords to emphasize the most important topics of their papers. Keywords can play an important role when other researchers use recommendation systems to discover works related to a specified set of input terms. In addition, journals provide keywords with the title and abstract as part of the public preliminary information about a paper, so this information can be crucial for a researcher to determine whether they will read the full paper or not. Nevertheless, keyword selection is a process that is not perfect and often has many problems, such as: selecting very specific or very generic keywords, misprints, author bias, inexperience, etc. These biases are aggravated by choosing keywords without following a keyword selection methodology.

When choosing generic or trendy terms, authors face the risk of sharing these terms with many papers that will compete with each other to claim researchers' attention. But in choosing rare or specific terms, we run the risk of using terms that do not attract researchers.

In this paper, we want to study how popularity relates to higher competition among keywords, when choosing keywords for a manuscript. We want to analyze ontologies, as doing so we would be able to intuitively understand whether generic terms

[1] CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0002-6334-3786.
[2] Corresponding autor.
[3] Departamento de Ciencias de la Computación e IA. CITIC-UGR. Universidad de Granada. 18071 Granada, Spain. ORCID: 0000-0001-7886-9329

tend to be more popular as well as more crowded than specific ones. We want to formalize some properties that will help us to study this phenomenon and we will continue by analyzing the structure of keywords by using an ontology.

Finally, after the analytic stage, we propose a method to help authors refine their keyword selection processes. This method is based on measuring the popularity and crowding of desired terms while using the knowledge provided by ontologies to enhance the keywords proposed by the author.

The paper is structured as follows:

1) Literature review: A description of the state-of-the-art and theory on which this paper is based.

2) The Attention-Survival model: A description of our theoretical proposal.

3) Experimental design: Information about the dataset we used, an analysis of the WordNet ontology and examples to illustrate our refinement algorithm.

4) Conclusions.

5) Bibliography.

6) Appendix.

Literature review

The International Organization for Standardization defines keywords as *a word or group of words, possibly in a lexicographically standardized form, taken out of a title or the text of a document characterizing its content and enabling its retrieval* (ISO 5963 1985). Apart from texts, keywords are often used to describe the content of a work by using words that contain the essential topics or themes that are represented in the work. For example, papers in the scientific community frequently come with a set of keywords, typically six. When authors decide what keywords they want to use for their manuscripts we call them Author Keywords. These keywords can be chosen freely by the author or selected from a pre-specified list of terms (Lu et al. 2020). Sometimes keywords are extracted by automatic procedures. That is the case for KeyWords Plus, where keywords are selected from the titles of articles cited in the references section (Zhang et al. 2016).

Keywords can have multiple applications with one of the most used being information retrieval, where the scientific community uses keywords to search for information about certain topics (Grant 2010; Hartley and Kostoff 2003; Sesagiri Raamkumar, Foo, and Pang 2017). Keywords are also used to easily identify the most relevant content of an article, study the behavior of authors (Gil-Leiva and Alonso-Arroyo 2007; González et al. 2018), map the structure of the science (Lozano et al. 2019), or build a taxonomy, among others (X. Liu et al. 2012).

Nowadays, plenty of search engines enable researchers to discover papers by typing in a set of keywords. Then, the search engine presents a list of recommended papers related to keywords, and the researcher can choose from them. This process has been analyzed by (H. Liu et al. 2020), who proposed a method to refine this recommendation process by choosing popular articles and very correlated keywords.

Aside from articles, keywords also have a different degree of popularity (Fernandes, Vinagre, and Cortez 2015). Keyword popularity is a topic that gained attention in the field of marketing research. (Jerath, Ma, and Park 2014) studied the different behaviors between customers who searched for popular terms and customers who used less popular keywords, concluding

that the second group of customers spent more effort on their search process and were more likely to buy something in the end.

It is also relevant to note that Author Keywords can prevent interpretation biases by other authors (González et al. 2018). Nevertheless, Author Keywords are not free from other types of biases, as there are differences between experienced and non-experienced authors (Sesagiri Raamkumar, Foo, and Pang 2017).

An author's behavior when choosing keywords was studied by (Hartley and Kostoff 2003). This paper performed a study on the habits of authors and editors regarding keywords, finding that, in the case of authors, it was very common to simply select as many keywords as desired. While, editors tend to let authors choose the keywords for a manuscript. (Hartley and Kostoff 2003) also focused on the problems generated by the inability of some authors to choose good keywords and the inefficiency of search systems. Some of the problems mentioned are the use of ambiguous keywords and the overuse of keywords without justification.

For authors, it is crucial to select correct keywords so that their papers are more easily visible to others. At the same time, from the editor's point of view, having a manuscript with proper keywords can help them improve their journal's impact (Pearce, Hicks, and Pierson 2018).

Some strategies have been proposed to mitigate the effects of selecting bad keywords. For example, (Zhang et al. 2016) grouped terms with a similar meaning into single primary terms while (Lozano et al. 2019), in addition to removing excessively specific words, divided generic terms into specific ones. Intuitively, it seems that using overly generic or overly specific keywords is bad practice.

In the Computer Science field, ontologies are an explicit specification of a conceptualization (Gruber 1993). Ontologies represent concepts and their relations in terms of generalization and specificity and can have different applications in fields such as artificial intelligence, web semantics, or linguistics (Dong et al. 2021; Guarino, Oberle, and Staab 2009).

The Attention-Survival model

Basic model

Our theoretical model is based on the premise that there is an information retrieval system where the user introduces a set of keywords. The system randomly returns a list of papers that contain all these keywords but in random order and from there the user chooses one of the returned articles. We consider that the paper selected by the user is the only one that 'survives' the process.

Our basic model does not consider different biases that would normally exist in an Information Retrieval System, such as ordering by citations, relevance, impact factor, publication date, etc., as well as the attention's bias generated by the user when choosing an article.

An example of a more complex model where the Information Retrieval System is biased with respect to the publication date is presented in the appendix.

The Attention-Survival score

Let $K_j = \{k_1, k_2, \ldots, k_n\}$ be the set of initial candidate keywords, as defined by the user for a manuscript j.

When authors search for manuscripts by entering specific inputs, we presume that only one manuscript will be extracted from each search process. Therefore, we denote that article selected as the survivor manuscript.

Let $C(k)$ be the community of a keyword k. We define community as the set of manuscripts that contains a given keyword. $C(k)$ defines the set of articles containing the keyword k and thus they are the ones that will compete with our manuscripts to survive.

Let $S(k), k \in K$ be the survival score of a keyword. Given an article which contains the keyword k, $S(k)$ is the propensity of the article to survive, according to our basic model. We assume that our recommendation system is neutral so that the retrieval process is unbiased. In order to help readers to understand our model, we are not considering values by relevance, length, cites, impact, or date of publication (basic model). If we apply a biased retrieval process, then we have to redefine $S(k), k \in K$ according to the bias applied. This situation is explained in *Biased models* Appendix. Under our basic model supposition, the survival score is defined as follows:

$$S(k) = \frac{1}{|C(k)|}$$

We consider that an author can look for either one keyword (e.g.,k) or a set of them (e.g.,K). When looking for multiple keywords at the same time, the community $C(K)$ is described as the set of manuscripts that contains every keyword in K simultaneously.

We can compute the survival score of K as follows:

$$S_{\cup}(K) = \frac{\sum S(k_i)}{|K|}$$

where $K = \{k_1, k_2, \ldots, k_n\}$.

Another important concept in our work is keyword attention,$A(k)$, which is the level of interest shown by the community for a certain keyword. As discussed later in the paper, the attention of a word is a function of the number of times that word is used in a query. We have derived this value with the information provided by Google Trends.

Similarly, we can compute keyword attention of a set as the average value of attention scores from every keyword in the input set. This is presented as:

$$A_{\cup}(K) = \frac{\sum A(k_i)}{|K|}$$

We define the attention-survival score,AS, as the score of a manuscript defined by K keywords. This score depends on the community and the attention of$K$.

$$AS_{\cup}(K) = \alpha \cdot S_{\cup}(K) + (1 - \alpha) \cdot A_{\cup}(K)$$

Where $\alpha \in [0,1], \alpha \in R$, is a weighting factor for $S_{\cup}(K)$ and $A_{\cup}(K)$

Finally, it is relevant to consider that, since Survival scores range from 0 to 1, the attention score will need to be normalized.

Keyword intersections

Sometimes, information retrieval systems search for the intersection of each term introduced by the user instead of treating each term separately. In that case, we need to adjust our expressions. Firstly, we introduce the survival of an intersection as follows:

$$S_{\cap}(K) = \prod S\,(k_i)$$

While the attention of an intersection is defined in the following way:

$$A_{\cap}(K) = \prod A\,(k_i)$$

Attention scores for each keyword should be computed or extracted from a reliable data source. Finally, we can adapt the attention-survival metric previously defined as follows:

$$AS_{\cap}(K) = \alpha \cdot S_{\cap}(K) + (1 - \alpha) \cdot A_{\cap}(K)$$

Theoretical behavior

Hypotheses and proposals

Hereafter, we will proceed by explaining the theoretical behavior of Survival and Attention among the different levels of an ontology: from the root node to the very last child on the tree. The intuitive idea behind our model is that the number of manuscripts that use a certain term tends to be higher when the community's interest in that term is also high.

Thus we propose the following hypotheses:

- **Hypothesis 1 (h1)**: The survival of a term is positively correlated with the distance of the term to the root node of the ontology.
- **Hypothesis 2 (h2)**: The attention of a term is negatively correlated with the distance of the term to the root node of the ontology.

Based on these hypotheses, we present three propositions as follows:

Proposition 1: The Survival score of a term depends on the specificity of that term inside the ontology structure so that the more specific a term is, the greater the Survival score will be.

Proposition 2: The Attention level of a term depends on the specificity of that term inside the ontology structure so that the more specific a term is, the less Attention it will achieve.

Proposition 3: For every term, there is a point where survival and attention intersect, and that is the equilibrium point.

It is trivial to state that, when the keyword does not have competitors the survival score tends to be ∞, $\lim_{C(k)\to 0} S(k) = \infty$as opposed to when we have infinite competitors the survival score would be zero, $\lim_{C(k)\to \infty} S(k) = 0$. In relation to the Attention score, if a term attracts the attention of infinite competitors the Attention will be at maximum, in contrast, when nobody is interested in that term the attention is zero.

Figure 3 graphically represents the *equilibrium point*. The equilibrium point is the level of specificity where the Attention and Survival scores intersect. Thus, the closest keyword to the equilibrium point would be the *equilibrium keyword*. The equilibrium point depends on various factors, such as α value and f1 and f2, which refer to the minimum level of keywords and the minimum interest in terms that can be used to find the maximum depth of an ontology.

So, if we choose a keyword that is more generic than the equilibrium keyword, we are reducing the survival score and with that the AS score will also decrease. While choosing a keyword that is less generic than the equilibrium keyword will reduce the AS score due to a reduction in attention.

Properties

We can use the idea behind the dynamics of Supply and Demand (Whelan, Msefer, and V.Chung 2001) from econometrics to better understand the behavior of the Attention and Survival functions. However, we must take into consideration that the behaviour tends to be slightly different between the two concepts:

Attention

Attention is a dynamic function that fluctuates over time, as it describes the behavior of people. Therefore, the popularity of a keyword is constantly changing. For example, Figure 1 illustrates the behaviour of the keyword “Support vector machine” vs. “Naive Bayes” from 2004 until 2021, according to Google Trends. As we can see, “Support Vector Machine” seems to be surpassed by “Naive Bayes” over time. Attention can play the same role as demand in economics theory. We also can interpret demand as the expected income a researcher hopes to receive when using a particular keyword.

A change in the popularity of a keyword produces a variation of the same sign in the attention function.

Survival

Survival is also dynamic and changes over time, as it describes the behavior of keywords. Survival can only grow to a fixed certain level based on the finite number of manuscripts that use a specific keyword. The role of survival might be similar to supply, but its dynamic is quite different. One approach can be to analyze Attention and Survival within a specific window of time so that survival would also be able to increase or decrease in response to tendencies. Survival can also be interpreted as the fixed price that a researcher must pay to use certain keywords.

A change in the number of manuscripts that contains a specific keyword produces a variation of an opposite sign in the survival function.

Complementary and substitutive keywords

Beyond the relationships so far discussed it should also be noted that keywords have relationships among themselves as well. In addition, sometimes people start using a new keyword to refer to an existing concept. Like in economic theory, a complementary keyword is a keyword whose popularity can, in turn, affect the popularity of the related keyword. At the same time, when a complementary keyword is affected in terms of survival, the complemented one is affected in the same way. This relationship is common in the case of synonyms, semantic parents or children, or keywords significantly correlated to one another. For example, the term “Machine Learning” is highly connected with the term “Artificial Intelligence”, as according to Google Trends (see Figure 2). With this information, we can see that complementary keywords have a positive correlation. When the complemented keyword gains popularity, the complementary keyword also increases in popularity. When the researcher community increases the use of one keyword, the other keyword also experiences an increase.

If one keyword completely replaces another one, then we are talking about substitutive keywords. When one substitute candidate keyword experiences an increase in attention, the attention received by the replaced keyword decreases. Similarly, when one keyword has a decrease in their survival score, the other one experiences a reduction in the rate that their survival decreases. If we use the window in time approach, there is a negative correlation between both keywords' survival score. An example of possible substitutive keywords are: “Support Vector Machine” vs. “Naive Bayes” (see Figure 1) or “C++” vs. “Python”.

As always, it is important to note that correlation does not imply causation. For example, both keywords “Digimon” and “Hip-hop” experienced a similar tendency on Google Trends, but there was no clear relationship between these two concepts. While

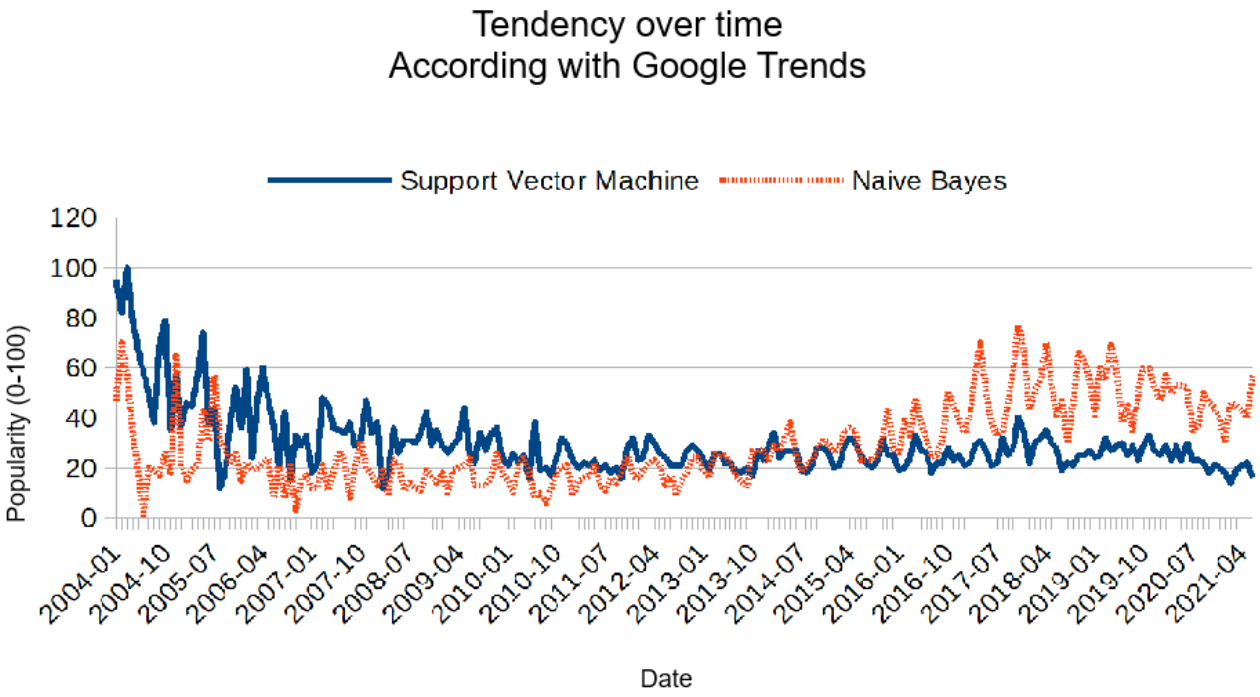

Figure 1: Global historic tendency of "Support Vector Machine" vs "Naive Bayes"

correlation can help us identify associations between keywords, we are required to further analyze the information to make decisive conclusions. For example, ontologies, lists of synonyms and antonyms, or analyses of social trends can help us to identify these associations.

Outsiders, Outlier Keywords, and Local Maximums

Not every keyword is part of an ontology relationship. For example, the “Me too” movement and the hashtag #MeToo have an important attention score (France 2017) and there are many academic manuscripts that use “MeToo” as a keyword (Blumell and Huemmer 2021). “Me too”, for example, is an outlier keyword if we are using WordNet, where this concept is not represented.

Often, children concepts have better attention or survival than their parents. For example, “AIDS” has a stronger popularity than their parent “immunodeficiency”, in WordNet. Even if these local maximums' existence is quite common within the ontology, the general tendency should follow the hypotheses posed in the previous section of our paper.

Candidate generation

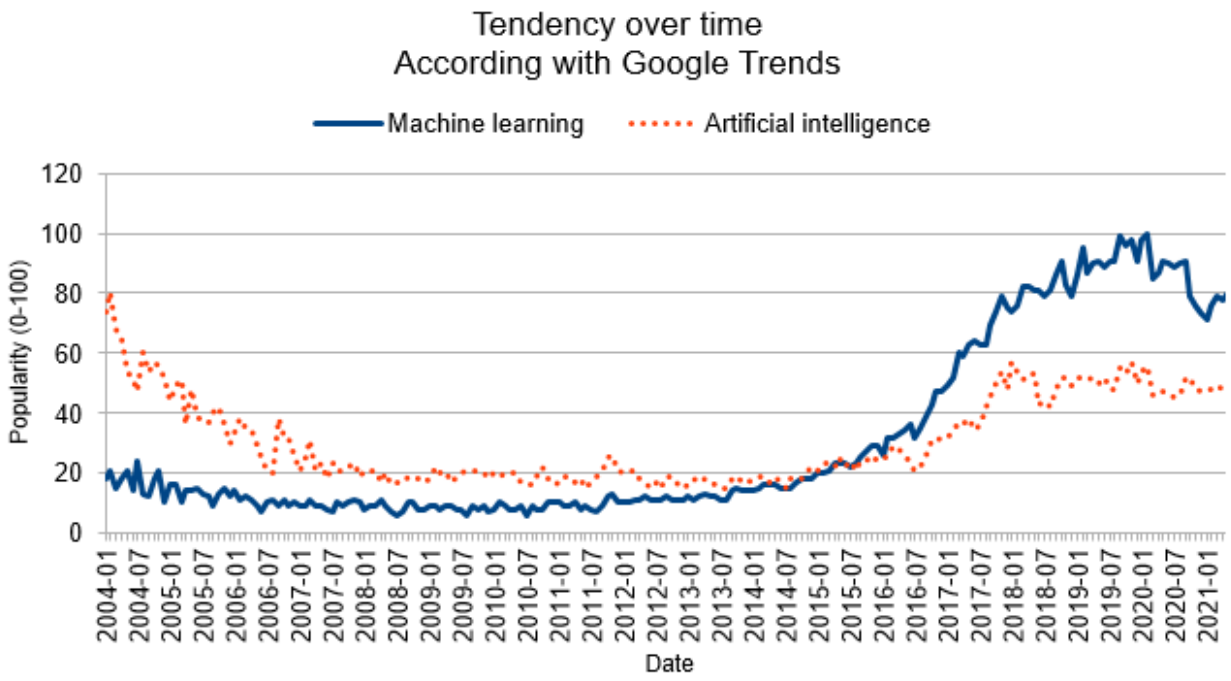

Figure 2: Global historic tendency of "Machine learning" vs "Artificial Intelligence".

We propose an iterative process to improve the Attention-Survival score of a manuscript's keywords by using their keyword 'neighbours'. To explore the neighborhood we can use a variety of techniques. In our paper, we propose using ontologies, as we can use human knowledge to determine the meaningful relationships of a keyword. Often, keywords have a very specific meaning, and it is important to change their semantic role as little as possible.

As ontologies are represented and defined as a tree, we must assume a trade-off between being general and being specific. By generalizing, we will often be able to increase the Attention score, but it will also increase the size of the community. Thus, an increase in $A(k)$ will often imply a decrease in $S(k)$. Conversely, moving to more specific keywords will increase $S(k)$ and decrease $A(k)$ as specific keywords are less searched than generic ones. For that reason, $\alpha$and 1-α play an essential role in the refining process.

It is important to take into consideration that ontologies can also contain synonyms (brother nodes). In relation to synonyms, it is difficult to predict what their effects on Survival and Attention would be.

Let $g(k_i, k_j) = \frac{1}{d(k_i,k_j)}$ to represent the benefit of selecting the keyword $k_i$instead of k_j in terms of distance between nodes. When $k_i = k_j$, g is 1. We define the evaluation function, $f(k_i, k_j, k_s)$as:

$$f(k_i, k_j, k_s) = AS(k_i) \cdot g(k_i, k_s) + AS(k_j) \cdot g(k_i, k_j)$$

where k_s is the starting candidate keyword.

Our goal here is to perform an iterative process to discover new candidate keywords and estimate whether paying the distance cost is worth increasing the AS score. Our iterative process is as follows:

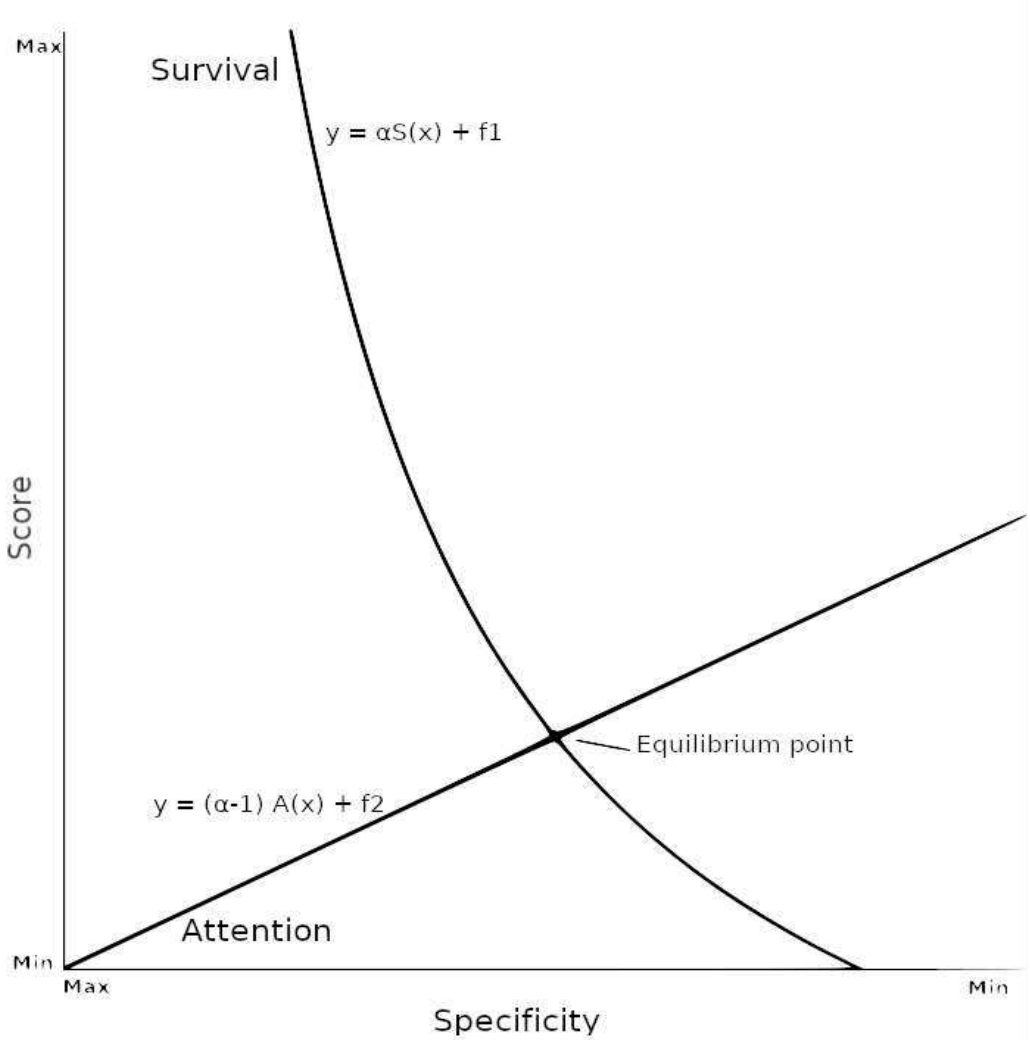

Figure 3: Graphical representation of an Equilibrium point.

Toy example

To illustrate our method, we have chosen a real paper that only has three keywords: “Chlorofluorocarbons”, “sorption” and “computer simulation” (George 1996) .

The first step is to choose one keyword for the queue iteration, for example, “Chlorofluorocarbons” and retrieve all their neighbours according to WordNet. Then, for each neighbor, we compute their Attention-Survival (AS) score as:

Fluorocarbon → 4,95; HCFC → 0,000264; Freon → 1,62; hydrochlorofluorocarbon → 0,40535; Chlorofluorocarbons → 0

1. We start with an initial candidate set , $K_{initial}$
2. $K_{candidates} = \{K_{initial}\}$
3. For each keyword, *k*, in $K_{initial}$:
   - 3.1. $k_{start} = k$ #k_{start} is the starting candidate keyword.
   - 3.2. $K_{successors} = \{k_{start}\}$ #*$K_{successors}$* is the set of potential replacements for k_{start}.
   - *3.3.* $K_{queue} = \{k_{start}\}$ #K_{queue} is a queue of candidates that have not been explored yet.
   - 3.4. While K_{queue} is not empty:
     - 3.4.1. $k_{source} = next(K_{queue})$ #gets the element out in front
     - 3.4.2. $K_{neighbours}$ is the list of neighbours of $k_{source,}$ including k_{source}.
     - 3.4.3. $AS_{neighbours}$ is the list of AS scores for each keyword in $K_{neighbours}$.
     - 3.4.4. $F_{neighbours} = \{f(ks_{ource}, k_j, k_{start})$ for each k_j in $K_{neighbours}\}$
     - 3.4.5. k_{best} is the keyword with the maximum *f* value in $F_{neighbours}$.
     - 3.4.6. If k_{best} is not in $K_{successors}$:
       - 3.4.6.1. $K_{successors} = K_{successors} \cup \{k_{best}\}$
       - 3.4.6.2. $K_{queue} = K_{queue} \cup \{k_{best}\}$
   - 3.5. For each keyword, $k_{candidate,}$ *in* $K_{successors}$ *:*
     - 3.5.1. For each candidate set, $K_{candidate\_set}$, in $K_{candidates}$:
       - 3.5.1.1. We create a new set, $K_{new\_candidate\_set}$ by replacing $k_{candidate}$ by *k*
       - 3.5.1.2. $K_{candidates} = K_{candidates} \cup K_{new\_candidate\_set}$
4. We return the set in $K_{candidates}$ that maximizes the AS score.

Algorithm 1: Keyword refinement algorithm

with “Fluorocarbon” being the neighbor with the best score. The next step is to compute *f*, which considers the gain by moving from the original term to one of their neighbours. In this example, we consider that the benefit of moving to a direct parent, children, or other brother terms in the ontology will always be the same distance, 1. For example, the *f* value of moving from “Chlorofluorocarbons” to their parent, “Fluorocarbon”, can be expressed as follows:

$$f(\text{chlorofluorocarbons}, fluorocarbon, chlorofluorocarbons) = AS(chlorofluorocarbons) \cdot$$
$$\cdot\, g(chlorofluorocarbons, chlorofluorocarbons) + +AS(fluorocarbon) \cdot g(chlorofluorocarbons, fluorocarbon)$$

With:

- $g(chlorofluorocarbons, chlorofluorocarbons) = 0$
- $g(chlorofluorocarbon, fluorocarbon) = 1$
- $AS(chlorofluorocarbons) = 0$
- $AS(fluorocarbon) = 4{,}95$

And afterwards, we compute f.

$$f(\text{chlorofluorocarbons}, fluorocarbon, chlorofluorocarbons) = 0 \cdot 1 + 4{,}95 \cdot 1 = 4.95$$

For the first step of the algorithm, getting the f values is trivial as the gain from moving to a direct neighbour is always one and k_i is the same as k_{start} so that all f values will ultimately coincide with their AS scores.

$$f(\text{chlorofluorocarbons}, HCFC, chlorofluorocarbons) = 0{,}002$$

$$f(\text{chlorofluorocarbons}, freon, chlorofluorocarbons) = 1{,}62$$

$$f(\text{chlorofluorocarbons}, hydrochlorofluorocarbon, chlorofluorocarbons) = 0{,}41$$

$$f(\text{chlorofluorocarbons}, chlorofluorocarbons, chlorofluorocarbons) = 0$$

“Fluorocarbon” was the best scored term, so we added “Fluorocarbon” to the candidate set as well as the queue for the following iteration.

We repeated the process, getting “Fluorocarbon” from the candidate set k_i = “Fluorocarbon”. . Note that our starting keyword in the algorithm, k_{start}, is “Chlorofluorocarbons”. When looking for neighbours and scores, we got 13 neighbours this time, with “Fluorocarbon” being the best of them. As “Fluorocarbon” was in the candidate set, we did not add it to the queue iteration.

The next step was to generate new candidate sets from the new candidate keywords. We proceeded by replacing the original keyword with the new candidate one. So that, our new list of candidates was:

{“Chlorofluorocarbons”, “sorption” and “computer simulation”}

{“Fluorocarbon”, “sorption” and “computer simulation”}

The next keyword to refine was sorption, which only had one good neighbour, “attention”. That meant we needed to add it to the new candidate sets:

{“Chlorofluorocarbons”, “sorption”, “computer simulation”}

{“Fluorocarbon”, “sorption”, “computer simulation”}

{“Chlorofluorocarbons”, “attention”, “computer simulation”}

{“Fluorocarbon”, “attention”, “computer simulation”}

Finally, it was time to refine “computer simulation”, which was a local maximum, meaning that this term did not have any neighbours with an AS score higher than its own AS score, so we did not add any new candidate sets. In conclusion, the best scored candidate set was **{“Fluorocarbons”, “attention”, “computer simulation”}**.

This entire process was firmly based on the knowledge from the ontology used (in our case, WordNet) and should be seen as a decision support system to help humans refine their keywords' impact. In the example, “Chlorofluorocarbons" was replaced by “Fluorocarbon”, which is a more generic concept that includes all “Chlorofluorocarbons". Authors must then judge whether it is worth the cost to accept the loss of specific information in order to use a more attractive keyword for the audience.

In the case of “attention”, “sorption” is the generic form of “absorption” so our algorithm moved to a child concept in order to finally end up with “attention”, which is another meaning of the keyword “absorption”. In the context of the article, it seems that "attention" is not a good choice to replace “sorption” because the manuscript context seems to be related to chemistry, not concentration. In that case, maybe the author would prefer to keep “sorption” or replace it with “absorption”, which is a bit more competitive term that receives more attention.

Experimental design

Theoretical model validation

Data source

To corroborate the validity of our hypotheses and have a better understanding regarding the behaviour of Attention and Survival on an ontology, we extracted data from a few different ontologies: WordNet (Miller 1995) and The Computer Science Ontology (CSO) (Salatino et al. 2018). WordNet is a lexical database that gathers words into groups of cognitive synonyms and defines relationships in terms of hypernymy and hyponymy. Thus, despite not being strictly an ontology, we can benefit from the WordNet structure, which also resembles the form of a tree where each concept is a node.

For their part, CSO is an ontology automatically generated from 16 million publications focused on the Computer Science field. The CSO model includes eight different semantic relations (relatedEquivalent, superTopicOf, contributsTo, preferntialEquivalent, rdf:type, owlSameAs, and schema:relatedLink). We only used the first two relations mentioned.

All terms from the ontologies are lightly pre-processed to avoid ambiguity and to prepare them to be sent to the APIs in a proper format. The pre-processing steps are the following:

1. Replace ‘-’ and ‘_’ with spaces.
2. Remove all characters except letters, spaces, and ‘&’.
3. Replace ‘&’ with ‘and’.

For extracting the number of papers according to Scopus, we used the Scopus Search API[4]. We performed requests to the Scopus Search API with the following filters:

- We looked for terms inside the keywords list of manuscripts. KEY(“term”).
- We filtered all the manuscripts except articles, reviews, and conference papers. DOCTYPE(“ar”), DOCTYPE(“re”), DOCTYPE(“cp”).

[4] Elsevier Developer Portal: https://dev.elsevier.com/

An example of a query search is as follows:

```
KEY ( 'SCIENCE' )  AND  ( LIMIT-TO ( DOCTYPE ,  'ar' )  OR  LIMIT-TO ( DOCTYPE ,  're' )  OR  LIMIT-TO ( DOCTYPE ,  'cp' ) )
```

From Google Trends, we extracted the popularity results per country and computed the average popularity.

Our purpose was to analyze the differences between a generic source of terms like WordNet and a more specific collection related to the Computer Science field. Scopus provided us with information about the number of papers per keyword necessary to infer Survival, while Google Trends gave us information regarding the attention and popularity of a term. Unfortunately, it is important to state that it was impossible to extract information for a specific academic search engine like Google Scholar. Instead, Google Trends gave us results for Google Search, which is used by the general population. Nevertheless, using results from Google can give us some additional information like altmetrics and the social interest in science topics.

Ontology analysis

Our purpose is to map the terms in our datasets onto the ontology structure as defined by WordNet and CSO. Before mapping terms, we first wanted to perform an exploratory task on both WordNet and CSO to explore both the ontologies' behavior and to check whether the theoretical process of Attention and Survival metrics over an ontology was close to our assumptions. As WordNet consists of different synsets outside a hierarchy, but we only studied those connected to the root synset (entity). A synset can have one or more lemmas, so we used the median value of the Attention and Survival scores across all lemmas. The use of the median value instead of the mean to determine the score of a synset is based on the fact that the distribution scores of lemmas tend to present a skewed result. However, we used the mean attention and survival numbers to compute the scores per level. Table 1 shows the distribution of Scopus and Google values over all levels of depths in WordNet ($\alpha = 0.5$). We only analyzed depth levels that were greater than five because for prior levels the number of synsets was too reduced, and that could lead to incorrect conclusions and inconsistent results.

After computing the Attention and Survival scores and extracting the mean values per level, we noted that the scores were in different scales, so we needed to perform a min-max normalization to keep all values in the same range [0,1]. We did this to make the analysis of the effect of both scores in the ontologies more accessible.

Figure 4 shows the Survival and Attention score evolution across the WordNet structure ($\alpha = 0.5$). As we can see, Survival starts close to 1.0 at depth 17 and is in a continuous decline until being surpassed by Attention at level 6, where the equilibrium point is located. Meanwhile, Attention is continually growing until it reaches its maximum value at level 2. The equilibrium point is located at the coordinates (6.38, 0.48), that is, in level 6, producing an AS score of 0.48 while the maximum score is achieved at level 2, with an AS score of 0.62.

If we check the results from Table 1, we can see that the average number of articles per level is significantly reduced until level 7, while Attention tends to grow uniformly.

The observed behavior over the WordNet ontology is in consonance with our proposed hypotheses (h1 and h2). Moreover, we have empirically corroborated that more specificity is related to low Attention and high Survival scores.

Concerning the WordNet dynamic, we can see how most depth levels contain very specific terms, which are quite unattractive according to Google searches and the number of articles retrieved by a Scopus search.

From these results, one should not deduce that the best option is to choose keywords from levels 7 or 2. WordNet is a generic ontology that contains many terms that are not common in the academic field. Therefore, a domain-specific ontology would be a better option to choose keywords from. A good approach could be to choose an ontology according to the criteria described by (Yu, Thom, and Tam 2007) (For example, the authors mention Clarity, Consistency, Conciseness, Expandability, Correctness, Completeness, Minimal Ontological Commitment and Minimal Encoding Bias, among other criteria). Our purpose in employing WordNet was to use a generic and widely validated ontology to analyze the distribution of Attention and Survival scores.

The case for CSO is illustrated in Figure 5.

In CSO, the equilibrium point was reached at level 10, while the maximum AS scored was in level 2. In CSO, survival fell quickly while attention had both fast-growing periods and periods of slow-growing. In CSO, the equilibrium point score was very close to the maximum value of AS (the difference was less than 0.02)

Both ontologies show a sudden drop on the first level. It is important to state that the first level is not the root node, which was removed from our data, but the upper levels of both ontologies contained such few words that their result could introduce noise into the graph and thus should be carefully interpreted.

Keyword refinement

In this section, we will randomly choose keywords from 20 manuscripts and we will run them through Algorithm 1. The Attention results will come from Google Trends and will be normalized with the interval [0-100] for this refinement process, thus, we do not need to perform a normalization step.
We limited the distance to the target keywords to two levels to prevent large differences in their conceptual meanings.

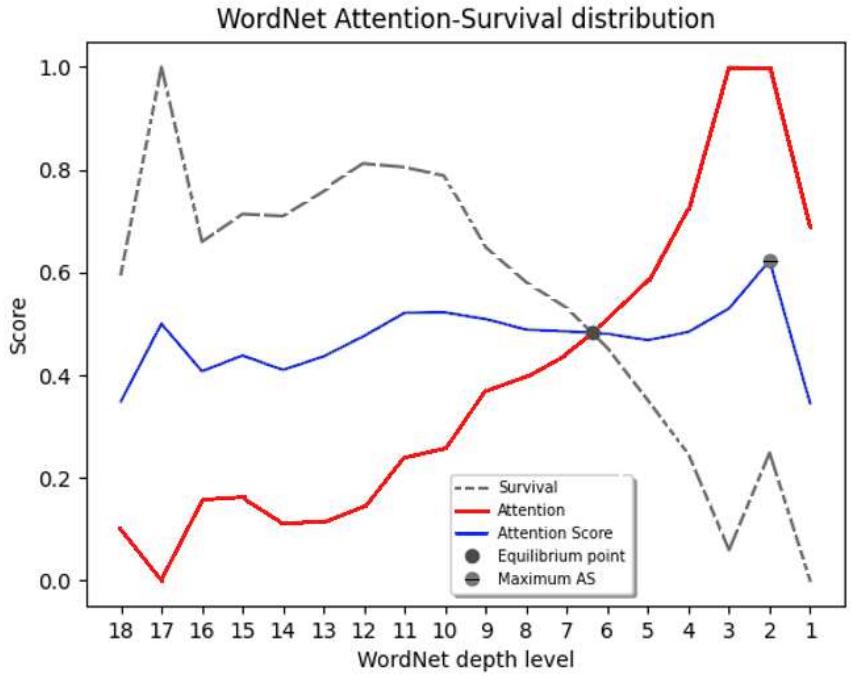

Figure 4: Evolution of Attention, Survival, and Attention-Survival Score (AS) throughout the WordNet structure. $\alpha = 0.5$

WordNet and CSO refinements

Table 2 shows some real examples of the author's keywords refined using WordNet and CSO ontologies[5]. Keywords were randomly selected from the intersection of terms contained in both ontologies. As CSO is generated from academic literature, all keywords in the example are real keywords. On the one hand, many keywords were not replaced by others. This can happen for two reasons: The algorithm is limited to explore at only a distance of two. If the neighbors' Attention-Survival score is low, the algorithm decides not to replace the keyword.
On the other hand, if the keyword is not defined in the ontology then we cannot use this ontology to refine the keyword as we do not have enough information about the neighborhood. For WordNet, the distance between terms is provided by the Path Distance Similarity, a metric that denotes how similar two synsets are based on the shortest path that connects the two nodes.

[5] An extended version of Table 2 can be found in the Supplemental Material of this paper.

This metric is provided by the Python Natural Language Toolkit (NLTK) library[6]. For CSO, we determined the distance between two words by using the Lowest Common Ancestor (LCA) algorithm (Aho, Hopcroft, and Ullman 1973).

Level	Scopus	Google
1	22.823,0	5,504
2	13.961,5	2,756
3	39.043,0	3,626
4	31.037,5	3,538
5	1.203,0	2,608
6	263,0	2,160
7	95,0	1,900
8	49,0	1,676
9	34,5	1,580
10	25,0	1,494
11	11,5	1,200
12	9,5	1,044
13	10,0	0,836
14	11,5	0,776
15	16,0	0,774
16	17,3	0,920
17	14,8	0,882
18	3,0	0,544
19	16,5	1,044

Table 1: Scopus and Google scores per depth level in WordNet. All values are the average scores of each synset with the same distance to the root node. And the value of each synset is computed by considering the median value of all lemmas within the synset. Note that level 1 only has the root synset.

[6] https://www.nltk.org/

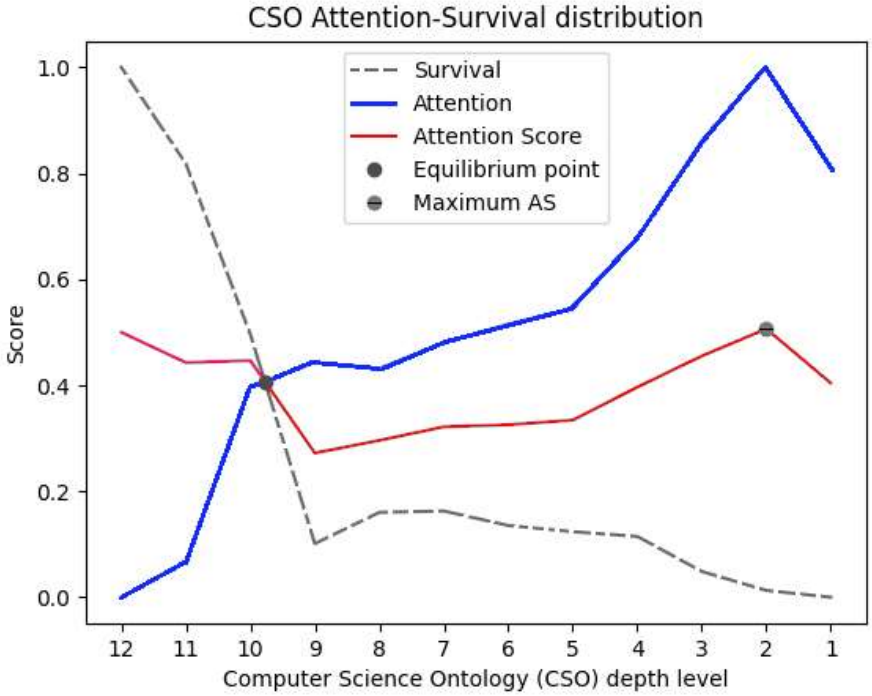

Figure 5: Evolution of Attention, Survival, and Attention-Survival Score (AS) throughout the CSO structure. $\alpha = 0.5$

As we mentioned before, WordNet is probably not the best option to use as an ontology, and a knowledge-specific ontology should be used instead (e.g., CSO for Computer Science or The Ontology for Biomedical Investigations for Biological or Medical domains (Bandrowski et al. 2016). We can see some replacements to keywords that perhaps are not the best option for manuscripts (Correspondence → card, Testing → Watch…).

The best way to use our method is inside an interactive system that allows the author to know the Survival and Attention scores from specific keywords and propose alternatives. Of course, the author should always make the final decision.

Initial word	WordNet Refinement	CSO Refinement
robotics	robotics	**robots**
telecommunication_equipment	**television**	**sensors**
electromagnetism	**acoustics**	**electromagnetic**
memory_access	memory_access	memory_access
computer-aided_design	**software**	**computer-aided**
gateway	gateway	**routing_protocols**
lexical_database	lexical_database	**artificial_intelligence**
speckle	speckle	**radar**
telecommunication_equipment	**television**	**sensors**
data_mining	**data_processing**	**clustering**
computer_science	**plan**	**software**
ergonomics	**technology**	**human_computer_interaction**
cosmic_microwave_background	cosmic_microwave_background	**polarimeter**
buffer_storage	**fund**	**bandwidth**
white_noise	**impediment**	white_noise
relational_database	relational_database	**database**
electrical_energy	**AC**	electrical_energy
mobile_phone	**cell**	**sensors**
binoculars	binoculars	**binocular**
object-oriented_programming	**hack**	**java**
user_interface	**CLI**	**sensors**
authentication	**validation**	**security_of_data**
remote_control	**device**	**robotics**
spline	**remove**	**computer-aided_design**

Table 2: List of different words before and after refining, using WordNet and CSO as ontologies[5].

WordNet and CSO Hierarchy

WordNet and CSO have different purposes as ontologies. While some terms are present in both ontologies, the knowledge structure is different between them. This can generate very different results from one ontology to another, as reflected in Table 2. In our paper, we also compared the hierarchy of both ontologies. Figure 6 and Figure 7 present the structure of the same set of keywords according to both the WordNet and CSO ontologies. As these terms are included in both ontologies, we can suppose that these keywords are closely related to the Computer Science field. Since CSO is an ontology focused on Computer Science terminology, we can see how these keywords are connected one to another and have less isolated nodes. For WordNet, however, most of these keywords are completely isolated, and there are not any strong clusters of keywords. Therefore, CSO represents terms with a higher granularity than WordNet and this situation directly impacts the refinement algorithm.

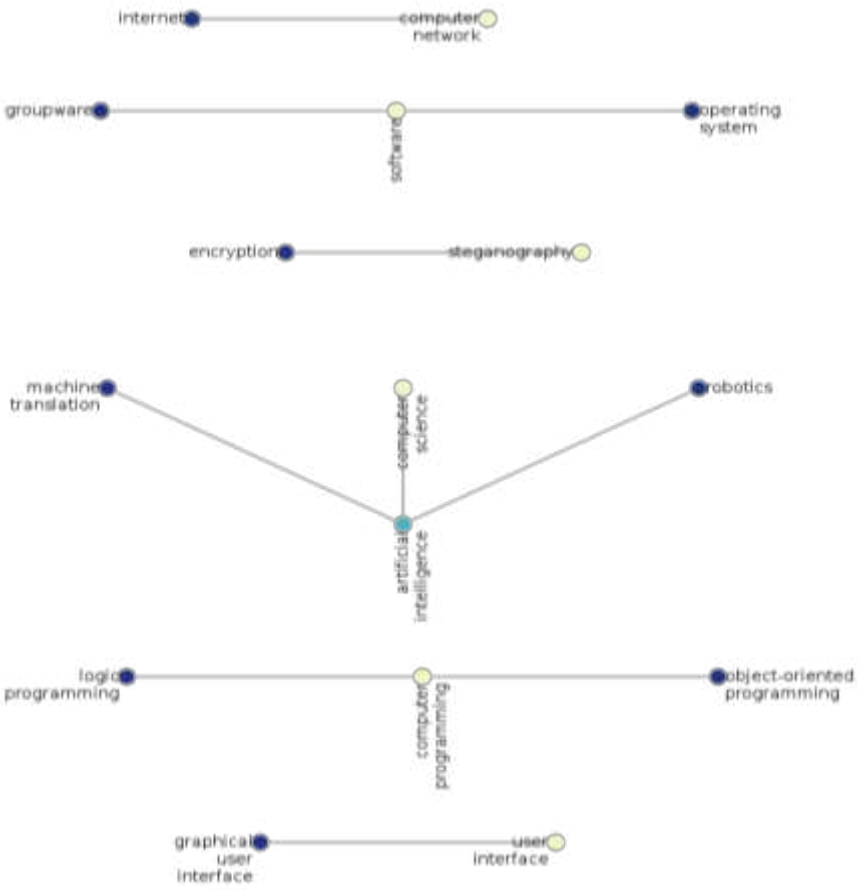

Figure 6: WordNet Hierarchy. The illustration represents the connections between different keywords in WordNet.

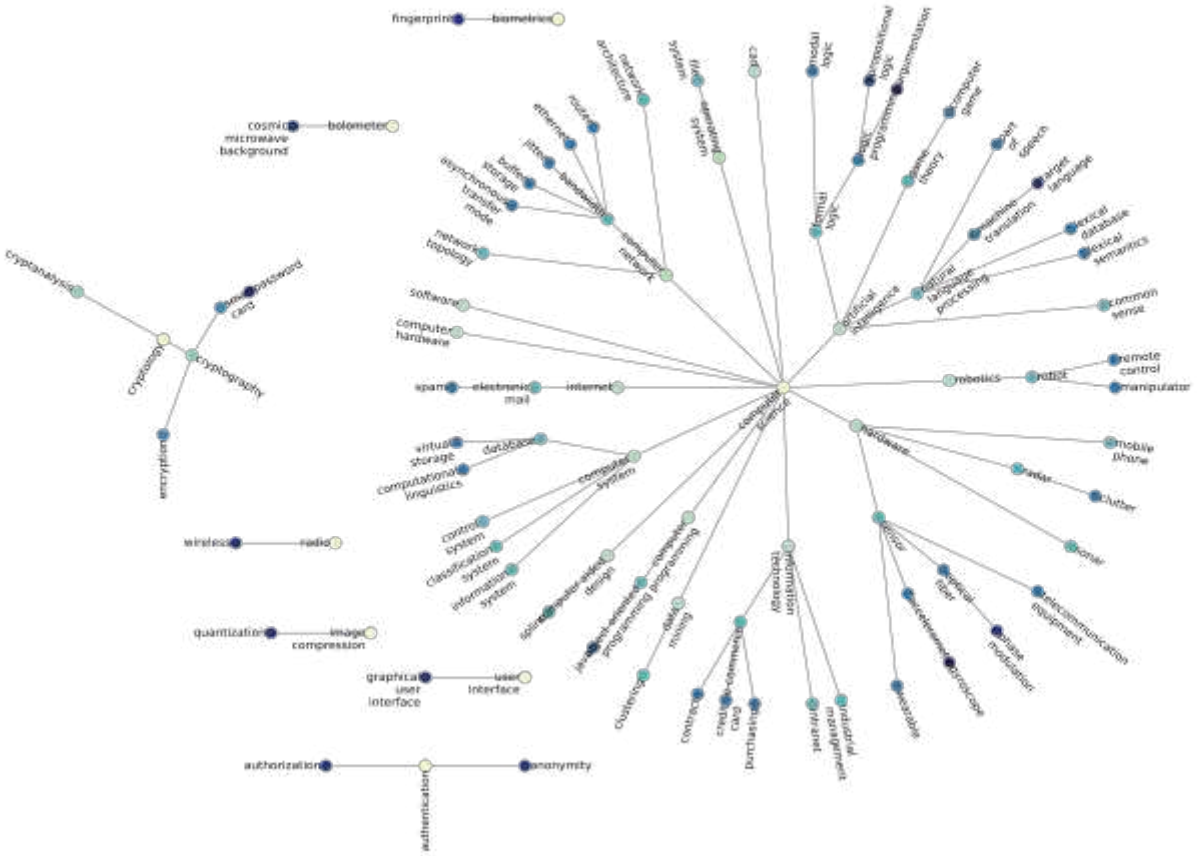

Figure 7: CSO Hierarchy. The illustration represents the connections between different keywords present in CSO.

Refinement algorithm validation

To validate the Algorithm 1, we have launched a survey in order to receive feedback from humans.

In first place, we extracted 500 articles from Web of Science by selecting articles from the category “Computer Science” and choosing only papers. Keywords from selected articles have been refined using the algorithm described in our work.

To deal with unknown terms, we firstly looked into CSO, which is the most related ontology with the Computer Science dataset we used. If the keyword is not present on CSO, then we try to find the concept in WordNet. When WordNet also fails, DBpedia[7] is used as a last chance. Finally, if the term is not presented in any ontology, we discarded the article and chose another one.

After having a set of refined articles, we radomnly chose 10 papers to be part of our survey. For each selected manuscript, we created a question in the survey where participants were allowed to see the title of the article and a table with two

[7] https://www.dbpedia.org

colums: one for the original set of keywords and another one for the refined set. The user had to select among one of the following answers:

- **R1**: The refined set can describe the title with almost the same precission than the initial set.
- **R2**: The refined set can't describe the title with almost the same precission than the initial set.

The supplemental material section contains all the questions asked in the survey.

Our survey were completed by 51 participants from Amazon Mechanical Turk. Participants were economically remunerated for their participation and had to meet the following preliminary requirements:

- Live in a English speaking country.
- A Bachelor degree.
- Working experience on IT.

With respect of survey results, for all questions, most of users answered R1. The Figure 9 describes the results obtained per question. The worst performance is obtained in q7, where 50.98% of participantes chose R1 while best results are obtained in q1 where 88.24% of participants checked R1. In the entire survey, the answer R1 was chosen by 67. 85% (standard deviation: 10.43).

Five participants chose R1 for all the questions while nobody selected R2 more times than R1. In global terms, participants chose R1 in 6.78 questions (standard deviation: 1.62).

As we can see on the Figure 10, the refined set of keywords deeply improve the AS score in comparison with the initial ones. The average improvement ratio is 22.5644 (standard deviation: 24.06).

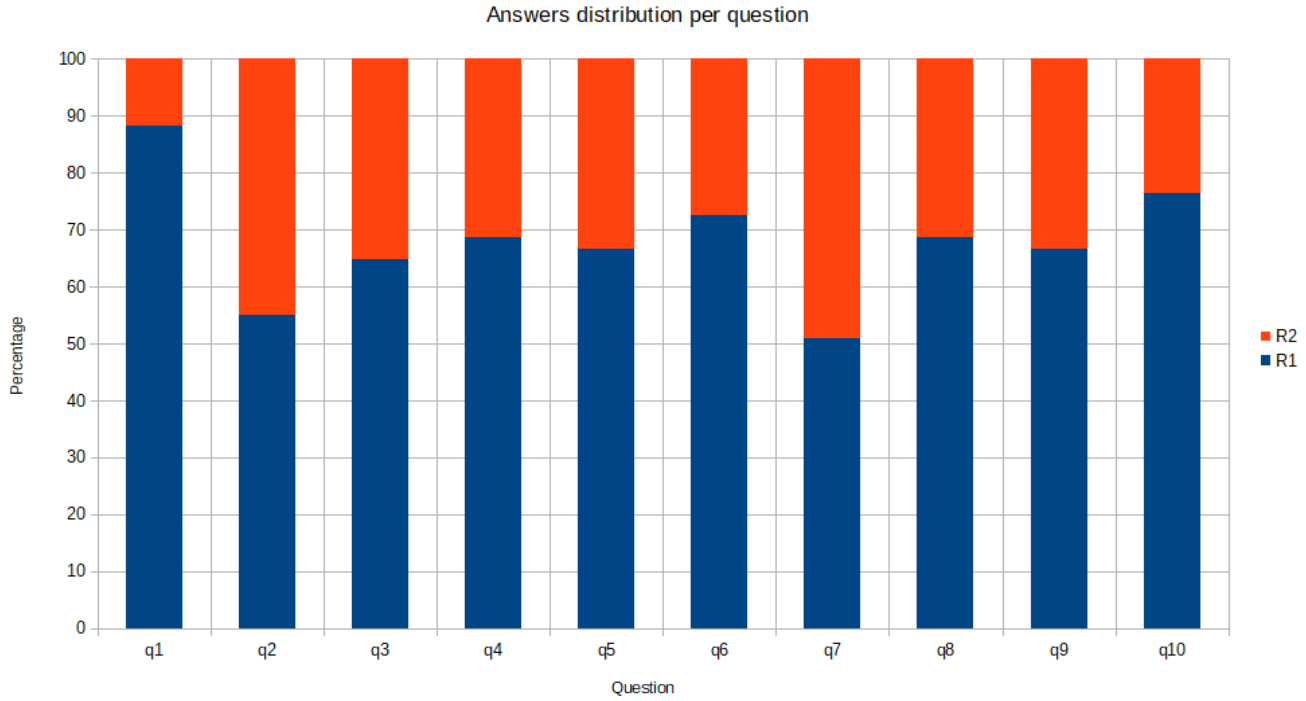

Figure 8: Answer distribution per question. Y-Axis illustrates the percentage of R1 and R2 per question.

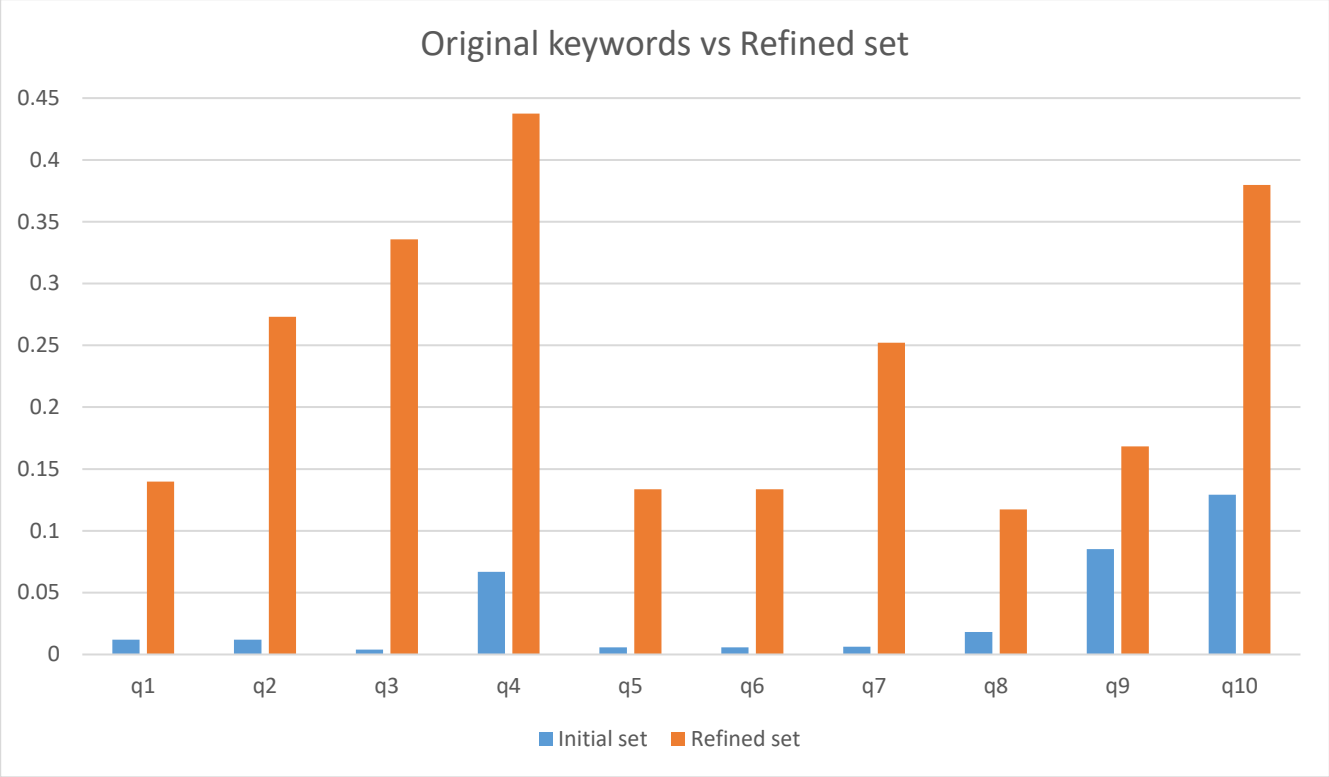

Figure 9:This figure compares the AS score obtained by the original set of keywords (blue bars) with the AS obtained by refined keywords (red bars).

Conclusions

Typically, keywords are selected by an authors' intuition or sometimes without applying any method at all. This can lead to bias, errors, and loss of opportunity.

The goal of our paper was to emphasize the importance of knowing the results for choosing different keywords. Choosing a keyword implies putting a future manuscript in competition with other ones, who all work to gain a certain amount of attention from the community that varies and depends on external factors. For that reason, keywords are constantly changing in terms of attention and survival rates. With respect to survival, we can say that all keywords decrease their survival possibilities as time goes by. But, in general terms, survival tends to decrease when moving from specific concepts to generic ones. At the same time, attention tends to decrease when moving from generic terms to specific ones. Sometimes, attention and survival intersect at certain equilibrium points. A keyword with both survival and attention scores that are simultaneously high, characterizes a keyword that will be used across time and will continue to be of interest to the community. To establish the survival and attention value of a keyword, we have defined the AS score.

We presented an algorithm to refine keywords by using ontologies in order to find alternatives keywords that have high survival and attention scores. Ontologies can be used as an essential source of knowledge that can help us organize keywords along the generic-specific axis. We analyzed WordNet and The Computer Science Ontology (CSO), both ontologies but with different backgrounds. CSO is a field-specific ontology, while WordNet gave us good comparison data to check how our model worked.

Implicitily, our method uses strategies defined in the State of the Art of our work to reduce the probability of choosing bad keyword (Zhang et al. 2016, Lozano et al. 2019) thanks to the implementation of ontologies and the possibility of moving into general or specific terms, according to the score obtained through the concepts hierarchy.

In our paper, we proposed two hypotheses: (**h1) "**The survival is positively correlated with the distance of a term to the root node of the ontology" and (**h2**) "The attention of a term is negatively correlated with the distance of a term to the root node of the ontology". According to the results from our experiments, both hypotheses are corroborated.

Another important topic is the human validation. We performed a survey where 51 participants answered positively to the results done by our algorithm, which used WordNet, CSO and DBpedia as ontologic sources.

In conclusion, it is important to state that our algorithm is intended to provide authors with additional information regarding how to choose keywords for a manuscript and propose some suggestions. However, the author is ultimately responsible for making the decision. Finally, our algorithm is not a keyword suggester, if an author makes a bad decision choosing a keyword, the refinement process likely will not help very much because it can only explore the related context of a keyword. In that case, the author must use some other methodology or information to select a good starting keyword candidate set.

Appendix

Biased models

The basic model is useful to help introduce our proposal and allows us to study the behavior of survival and attention scores without having any bias. Nowadays, Academic Information Retrieval Systems usually tend to return results using a concrete aspect of the manuscript (date of publication, impact, number of citations, journal, relevance, altmetrics, etc.), so there are plenty of biases to take into consideration to estimate survival and attention better.

A straightforward case of biased Information Retrieval Systems is a system that keeps a prioritized list of papers according to a certain parameter (date, relevance, etc.). The list is ordered in descending order of survival scores so that the first paper in a prioritized list of n papers has n times more survival score than the last document in the list. In this situation, we need to redefine survival to consider the position of the paper in the list. The biased survival score of a manuscript, S_{biased}, could be expressed as follows:

$$S_{biased}(p_t, K) = \frac{|\{p: Pos(p,K) \leqslant Pos(p_t,K)\}| / p \in C(K)}{\sum_{i=1}^{|C(K)|} i}$$

where p_t is the target paper of whose survival score we want to study. K is the set of keywords introduced by the user and $Pos(p,K)$ is the position of the paper in the biased list returned by the Information Retrieval System so that the first paper returned by looking for papers that contains the keywords in the set K is p_1.

Bibliography

Aho, A. V., J. E. Hopcroft, and J. D. Ullman. 1973. "On Finding Lowest Commong Ancestors in Trees." In *Http://Dx.Doi.Org/10.1137/0205011*, ACM, 253–65.

Bandrowski, Anita et al. 2016. "The Ontology for Biomedical Investigations." *PLoS ONE* 11(4).

Blumell, Lindsey E., and Jennifer Huemmer. 2021. "Reassessing Balance: News Coverage of Donald Trump's Access Hollywood Scandal before and during #metoo." *Journalism* 22(4): 937–55. http://journals.sagepub.com/doi/10.1177/1464884918821522 (June 5, 2021).

Dong, Sicong, Yike Yang, He Ren, and Chu-Ren Huang. 2021. "Directionality of Atmospheric Water in Chinese: A Lexical Semantic Study Based on Linguistic Ontology." *SAGE Open* 11(1): 215824402098829. http://journals.sagepub.com/doi/10.1177/2158244020988293 (May 9, 2021).

Fernandes, Kelwin, Pedro Vinagre, and Paulo Cortez. 2015. "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 535–46.

France, Lisa Respers. 2017. "#MeToo: Social Media Flooded with Personal Stories of Assault." *CNN*. https://web.archive.org/web/20171016002502/http://www.cnn.com/2017/10/15/entertainment/me-too-twitter-alyssa-milano/index.html (June 5, 2021).

George, A. R. 1996. "A Computational Investigation of Zeolite-Chlorofluorocarbon Interactions." *Zeolites* 17(5–6): 466–72.

Gil-Leiva, Isidoro, and Adolfo Alonso-Arroyo. 2007. "Keywords given by Authors of Scientific Articles in Database Descriptors." *Journal of the American Society for Information Science and Technology* 58(8): 1175–87. http://doi.wiley.com/10.1002/asi.20595 (January 15, 2020).

González, Luis Millán et al. 2018. "An Author Keyword Analysis for Mapping Sport Sciences." *PLoS ONE* 13(8): 1–22.

Grant, Maria J. 2010. “Key Words and Their Role in Information Retrieval.” *Health Information & Libraries Journal* 27(3): 173–75. http://doi.wiley.com/10.1111/j.1471-1842.2010.00904.x (May 7, 2021).

Gruber, Thomas R. 1993. “A Translation Approach to Portable Ontology Specifications.” *Knowledge Acquisition* 5(2): 199–220. https://linkinghub.elsevier.com/retrieve/pii/S1042814383710083 (May 9, 2021).

Guarino, Nicola, Daniel Oberle, and Steffen Staab. 2009. “What Is an Ontology?” In *Handbook on Ontologies*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1–17. http://link.springer.com/10.1007/978-3-540-92673-3_0 (May 9, 2021).

Hartley, James, and Ronald N. Kostoff. 2003. “How Useful Are ‘key Words’ in Scientific Journals?” *Journal of Information Science* 29(5): 433–38. http://journals.sagepub.com/doi/10.1177/01655515030295008 (May 7, 2021).

ISO 5963. 1985. “ISO/IEC 5963:1985 Documentation - Methods for Examining Documents , Determining Their Subjects , and Selecting Indexing Terms.” *Iso 5963:1985*: 3–5. https://www.iso.org/standard/12158.html.

Jerath, Kinshuk, Liye Ma, and Young-Hoon Park. 2014. “Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity.” *Journal of Marketing Research* 51(4): 480–86. http://journals.sagepub.com/doi/10.1509/jmr.13.0099 (May 9, 2021).

Liu, Hanwen, Huaizhen Kou, Chao Yan, and Lianyong Qi. 2020. “Keywords-Driven and Popularity-Aware Paper Recommendation Based on Undirected Paper Citation Graph.” *Complexity* 2020.

Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. “Automatic Taxonomy Construction from Keywords.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA: ACM Press, 1433–41. http://dl.acm.org/citation.cfm?doid=2339530.2339754 (May 7, 2021).

Lozano, S., L. Calzada-Infante, B. Adenso-Díaz, and S. García. 2019. “Complex Network Analysis of Keywords Co-Occurrence in the Recent Efficiency Analysis Literature.” *Scientometrics* 120(2): 609–29. https://doi.org/10.1007/s11192-019-03132-w.

Lu, Wei et al. 2020. “How Do Authors Select Keywords? A Preliminary Study of Author Keyword Selection Behavior.” *Journal of Informetrics* 14(4): 101066.

Miller, George A. 1995. “WordNet: A Lexical Database for English.” *Communications of the ACM* 38(11): 39–41. http://portal.acm.org/citation.cfm?doid=219717.219748 (October 23, 2018).

Pearce, Patricia F., Rodney W. Hicks, and Charon A. Pierson. 2018. “Keywords Matter: A Critical Factor in Getting Published Work Discovered.” *Journal of the American Association of Nurse Practitioners* 30(4): 179–81.

Salatino, Angelo A. et al. 2018. “The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 187–205.

Sesagiri Raamkumar, Aravind, Schubert Foo, and Natalie Pang. 2017. “Using Author-Specified Keywords in Building an Initial Reading List of Research Papers in Scientific Paper Retrieval and Recommender Systems.” *Information Processing and Management* 53(3): 577–94.

Whelan, Joseph, Kamil Msefer, and Celeest V.Chung. 2001. *Economic Supply & Demand*. Cambridge, Mass. : MIT, 2001.

Yu, Jonathan, James A. Thom, and Audrey Tam. 2007. “Ontology Evaluation Using Wikipedia Categories for Browsing.” In *International Conference on Information and Knowledge Management, Proceedings*, New York, New York, USA: ACM Press, 223–32. http://portal.acm.org/citation.cfm?doid=1321440.1321474 (June 3, 2021).

Zhang, Juan et al. 2016. “Comparing Keywords plus of WOS and Author Keywords: A Case Study of Patient Adherence Research.” *Journal of the Association for Information Science and Technology* 67(4): 967–72.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- finalversionsupplemental.docx