

# Representing, Tracking and Revising the User's Knowledge: A Search Result Filter Framework

Discussion Paper

Dima El-Zein, Célia da-Costa-Pereira

## Abstract

This paper presents a framework for a cognitive agent in information retrieval that personalizes the list of returned documents based on what it believes about the user's knowledge. Throughout the interactions between the IR system and the user, the agent builds its beliefs about the user's knowledge by extracting keywords from the content of the documents read by the user. The agent's belief base, which corresponds to the user model, contains also "contextual rules" that allow deriving new beliefs about the user's knowledge. The agent is therefore able to compare its own beliefs with the content conveyed by a to-be-proposed document, and thus understand if the document really contains useful information for the user or not. Finally, in case of beliefs' inconsistency, the agent revises its belief base to restore consistency.

## Keywords

Search Filter, Information Retrieval, Cognitive Agent, Knowledge Extraction, Belief Revision

## 1. Introduction

In the domain of information retrieval, it is not always sufficient to return the information responding only to the query. It is believed that users can be considered as cognitive agents having their own beliefs and knowledge about the world [1]. They try to fulfill needs for information by requesting queries and acquire new information by examining the results. In consequence, the search results must also respond to the user beliefs, knowledge, and search goals. Considering the user's cognitive components in the domain of Information Retrieval has been set as one of the "major challenges" by the IR community in 2018 [2].

In this paper, we propose an Information Retrieval filter framework that uses the content of the documents read by the user to learn about his/her knowledge. This cognitive awareness is employed to personalize the returned documents with respect to what the user already knows. To our knowledge, there are no research dealing with the content of the documents read by the user as his/her acquired knowledge.

The framework we have proposed in [3] works as follows. For every submitted query: (i) the system sends the user's query to the search engine and receives a list of documents relevant to the query (ii) the agent examines the content of the documents in the list and measures the similarity between each document and the set of beliefs (iii) the agent returns a filtered list according to the similarity results (iv) the user reads a proposed document (v) the agent adds the


---

*IIR 2021 – 11th Italian Information Retrieval Workshop, September 13–15, 2021, Bari, Italy*

 [elzein@i3s.unice.fr](mailto:elzein@i3s.unice.fr) (D. El-Zein); [Celia.DA-COSTA-PEREIRA@univ-cotedazur.fr](mailto:Celia.DA-COSTA-PEREIRA@univ-cotedazur.fr) (C. da-Costa-Pereira)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

keywords representing the read document as new beliefs (vi) A reasoning cycle is performed to derive new beliefs and revise the belief base if needed.

## 2. Proposed Framework

The agent is modeled as a *Rule-based* agent that consists of beliefs (ground literals) and rules (Horn clauses). The beliefs represent what the agent believes about its user's knowledge. When the agent has  $\alpha$  in its belief base, it believes that the user knows that  $\alpha$  is true. If the belief base contains  $\neg\alpha$ , then the agent believes the user knows that  $\alpha$  is not true. On the other side, rules are the relationships between beliefs, that are used to derive new beliefs from the agent's existing ones. Rules have the form of  $\alpha_1 \& \alpha_2 \dots \& \alpha_n \rightarrow \beta$  where  $\alpha_1, \alpha_2, \dots, \alpha_n$  ( $n \geq 1$ ) and  $\beta$  are literals.  $\beta$  is called the *derived belief*, and each belief  $\alpha_i$  is a premise of the rule. The  $\&$  symbol represents the logical *and* operator. During an agent's reasoning cycle, if all the premises of the rule are satisfied (the premise exists in the belief base), the rule is fired and  $\beta$  is added to the belief base. The rules will be considered static, their extraction/origin will not be discussed in this paper.

The agent acquires its beliefs about the user's knowledge from the documents the user has read. When the user reads a document  $d$ , the agent extracts the content of the document and considers it as an acquired knowledge by the user. We propose applying the RAKE - Rapid Automatic Keyword Extraction Algorithm - [4] as an easy and understandable method, to extract the set of scored keywords representing the document. Those keywords will be associated with the agent's extracted beliefs. The belief base components are then both, the extracted and derived beliefs, and the rules.

We consider that a belief is gradual and an agent might have beliefs more entrenched (or accepted) than others. We define a "degree" for beliefs to measure this entrenchment:

**Definition 1.** *The degree of a belief  $\alpha$  is the degree to which the agent believes the user is knowledgeable about  $\alpha$ . It is represented by a decimal ranging between 0 and 1, where 0 means the lowest degree –the agent believes the user has absolutely no knowledge about  $\alpha$ , and 1 means the highest degree –the agent believes the user has the maximum knowledge about  $\alpha$ .*

*Let us define document  $d = \{(k_1; s_1) \dots, (k_n; s_n)\}$  as a set of tuples where  $k_i$  is the keyword extracted by RAKE and  $s_i$  is its related score;  $k_i$  will be associated with an **extracted belief**  $b_j$  whose degree is calculated as follows:*

$$degree(b_j) = \lambda \cdot \frac{s_i}{\max_{s_j \in d}(s_j)}. \quad (1)$$

In Equation 1 the RAKE score of an extracted keyword is normalized then multiplied by an adjustment factor  $\lambda \in [0, 1]$  that weakens the magnitude of the degrees. The adjustment factor may vary based on different characteristics such as the trust on the document's source, for example.

This equation allows the calculation of the degree for *extracted beliefs* only. As for *derived beliefs*, their degrees will depend on the degree of premises that derived them. For that reason, we track the dependency between the beliefs by following the approach proposed by Alechina

et al. [5] in which the dependency between beliefs is tracked as follows. For every fired rule instance, a *Justification*  $J$  will record: (i) the derived belief and (ii) a *support list*,  $s$ , which contains the premises of the rules. The dependency information of a belief has the form of two lists: *dependencies* and *justifications*. A *dependencies list* records the justifications of a belief, and a *justifications list* contains all the Justifications where the belief is a member of a support.

The degree value of a *derived belief*  $b$ ,  $degree(b)$ , is equal to that of its highest quality justification.

**Definition 2.**

$$degree(b) = \max\{qual(J_0), \dots, qual(J_n)\} \quad (2)$$

**Definition 3.** The quality of justification  $J$ ,  $qual(J)$ , is equal to the degree of the least entrenched belief in its support list.

$$qual(J) = \min\{degree(b) : b \in \text{support of } J\} \quad (3)$$

For example, suppose an agent has two beliefs *planets* and *stars* with degree equal to 0.5 and 0.7 respectively. The belief base has also a rule *planets & stars*  $\rightarrow$  *galaxies*. It means that if the agent “believes” in stars and planets, it will believe in galaxies. When the rule is fired, a Justification  $J_1$  denoted as  $(galaxies, [planets, stars])$  will be added; *galaxies* is the derived belief and  $[planets, stars]$  is the *support list*. The *quality* of  $J_1$  is equal to  $\min\{degree(planets); degree(stars)\} = 0.7$ .  $J_1$  is in the *dependencies* list of *galaxies* and in the *justifications* list both *planets* and *stars*.

While the agent is acquiring more information about the user, it is adding more beliefs to its belief base. The beliefs might be new, already existing, or contradicting with the existing ones; that calls for the need of revising beliefs to ensure the belief base is consistent.

Belief revision is the process of modifying the belief base to maintain its consistency whenever new information becomes available. We follow the AGM belief revision theory [6] that defines postulates a rational agent should satisfy when performing belief revision. We consider a belief base  $K$  and a new piece of information  $\alpha$ .  $K$  is inconsistent, when both  $\alpha$  and  $\neg\alpha$  are in  $Cn(K)$ , or  $Cn(K) = \perp$ , or both  $\alpha$  and  $\neg\alpha$  are logical consequences of  $K$ . Three operators are considered: *Expansion*  $K + \alpha$ : adds a new belief  $\alpha$  that does not contradict with the existing beliefs. *Contraction*  $K \div \alpha$ : removes a belief  $\alpha$  and all other beliefs that logically imply/entail it. *Revision*  $K * \alpha$ : adds a belief  $\alpha$  as long as it does not cause a contradiction in  $K$ .

In our framework, if the addition of a belief  $\alpha$  will cause inconsistencies in  $K$  (because of the existence of a  $\neg\alpha$ ), the priority/preference is given to the belief with the higher degree: In case  $\alpha$  has the higher degree, the revision operation starts with minimal changes in  $K$  to make it consistent with  $\alpha$ , contracts  $\neg\alpha$ , then adds  $\alpha$ . If  $\neg\alpha$  was a derived belief, we do not contract other beliefs that derived  $\neg\alpha$ , as long they are consistent with the remaining beliefs (*minimal change*) – *coherence approach* [7]. In other words, we only contract the belief in question with its related justification(s), without contracting neither the rule’s premises nor the rule itself. In case  $\neg\alpha$  had the higher degree, then the addition of  $\alpha$  is discarded.

The filtering process is based on the similarity  $Sim(B, d)$  between the agent’s set of beliefs  $B = \{(b_1; degree(b_1)), \dots, (b_m; degree(b_m))\}$  and the content of a proposed document  $d = \{(k_1; s_1), \dots, (k_n; s_n)\}$  to be proposed to the user. We propose a similarity measure that

considers the degrees of the intersected beliefs and the knowledge in the document. The formula is inspired by the similarity function proposed by Lau *et al.* in [8]. Let us consider  $S$ , the set of keywords appearing both in  $d$  and in  $B$  defined by  $S = \{k_i \in d : extent(B, k_i) > 0 \vee extent(B, \neg k_i) > 0\}$ .

$$Sim(B, d) = \begin{cases} \frac{\max\{\sum_{k_i \in d} [extent(B, k_i) - extent(B, \neg k_i)], 0\}}{|S|} & \text{if } |S| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

The  $extent(B, k_i) = degree(k_i)$ , if  $k_i \in B$ ; and 0 otherwise. The similarity formula “rewards” the documents containing common keywords with the set  $B$  and penalizes those containing keywords whose corresponding negated beliefs are in  $B$ .

We set a cutoff value  $\gamma$  for  $Sim(B, d)$  that allows to decide whether the knowledge inside a document is similar to a set of beliefs or not. The filter is used according to the intended application: when the purpose of framework is reinforcing the user’s knowledge, then documents that are “close” to the agent’s beliefs will be returned. The documents having a similarity score greater than the cutoff will be returned to the user. Contrarily, when the framework is employed for novelty, the documents having similarity below the cutoff will be returned.

### 3. Conclusion

This paper proposed an innovative framework for a rule-based information retrieval agent which relies on its cognitive abilities to learn about the user’s knowledge. This information is used to propose new/relevant documents accordingly. The components of the framework are: (1) Rule-based module, modeling the agent’s beliefs and rules. It performs inference reasoning about the user’s knowledge, calculates the entrenchment degrees and tracks the dependency between them. It also revises the beliefs if needed, to maintain consistency. (2) Knowledge extractor module, extracting knowledge from the documents read by the user. (3) Result filtering module, compares the content of the potential to-be-proposed documents to the user’s knowledge and select the “useful” ones to be returned to the user.

For future work, we aim to take into account the confidence in the sources of the documents, which will probably affect the degree of entrenchment of a belief. Another possible extension could be to integrate some semantic analysis to deal with semantically similar content.

### References

- [1] M. da Costa Móra, J. G. P. Lopes, R. M. Vicari, H. Coelho, Bdi models and systems: Bridging the gap., in: ATAL, 1998, pp. 11–27.
- [2] J. S. Culpepper, F. Diaz, M. D. Smucker, Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018), SIGIR Forum 52 (2018) 34–90.
- [3] D. El Zein, C. da Costa Pereira, A cognitive agent framework in information retrieval: Using user beliefs to customize results, in: The 23rd International Conference on Principles and Practice of Multi-Agent Systems, 2020.

- [4] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, *Text mining: applications and theory 1* (2010) 1–20.
- [5] N. Alechina, M. Jago, B. Logan, Preference-based belief revision for rule-based agents, *Synthese* 165 (2008) 159–177.
- [6] C. E. Alchourrón, P. Gärdenfors, D. Makinson, On the logic of theory change: Partial meet contraction and revision functions, *The journal of symbolic logic* 50 (1985) 510–530.
- [7] P. Gärdenfors, *Belief revision: An introduction*, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 1992, pp. 1–28.
- [8] R. Y. Lau, P. D. Bruza, D. Song, Belief revision for adaptive information retrieval, in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 130–137.