5
Evaluation

In this section, we analyze the proposed search system approach.

There are some standards measures to evaluate the performance of IR systems. To evaluate correctness (task-performance) and effectiveness (time-performance) of a method of retrieval software artifacts in the repository, we rely on two metrics that have traditionally been used to, recall and precision. Recall means to get all the relevant components. Precision means that all the retrieved components are exact as per query submitted by a user.

- *Recall*: The ratio of artifacts retrieved by the system that are actually relevant to the query divided by the total number of relevant artifacts in the repository.

  The synonymy leads to lower recall rates because relevant agents referring only to synonyms of a word used in the query may not be actually retrieved.

- *Precision*: The ratio of artifacts retrieved by the system that are actually relevant to the query divided by the total number of artifacts retrieved.

  For instance, if the system retrieves 4 agents for a query, where 2 of them are really relevant, the performance of the precision for the system in that query has value of 0.5 The polysemy may produce low precision rates due to irrelevant agents might be retrieved.

Ideally, a search mechanism must have good precision and good recall. High recall means that few relevant elements are left behind, without being retrieved. High precision means that most retrieved elements are relevant.

To evaluate our information model, all agent components already stored in the repository, which belong to different application domains, and a produced test collection are involved. A test collection is a set of queries elaborated that cover all the components and the associated set of relevant components that is known a

priori. This is possible due to our repository does not contain a lot of agent-oriented-artifacts.

Then, the search of each query is performed using mainly the keyword-based, tag-based and interface-based approaches, with a fixed minimum relevance acceptance value of 0.6. We picked the value 0.6 as threshold according to [50]. How we know a priori which agents are really relevant among all that are returned to that query and how many relevant artifacts are in the repository, we can calculate the recall. Additionally we know the quantity of agents retrieved so, we can calculate the precision too.

We take into consideration the three different types of searches we implemented within the repository. We realize that the values of recall and precision for these techniques have the same behavior in the case of the searches based on language, platform, categories and tags. Just in this particular case, an exact match of the query and the respective attribute was made. But, for interface-based and keyword-based searches, the system behaves different how we supposed.

Table 2 shows the results of recall (R) and precision (P) for specific queries in the three different search methods.

Table 2: Evaluation Results.

Approach / Queries	Keyword	SPARQL	Lucene+ WordNet
agent simulation	R: 0.4 P: 0.4	R: 0.9 P: 1.0	R: 1.0 P: 1.0
didactic games	R: 0 P: 0	R: 0.666 P: 0.333	R: 1.0 P: 0.9166
robotics	R: 0 P: 0	R: 0.8 P: 0.777	R: 1.0 P: 1.0
negotiation books	R: 0 P: 0	R: 0.5 P: 0.6	R: 0.6666 P: 0.857

To conclude, we find that semantics can significantly improve precision and recall of search approaches.