Journal of Theoretical and Applied Information Technology
15th April 2019. Vol.97. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE ONTOLOGY APPROACH FOR INFORMATION RETRIEVAL IN LEARNING DOCUMENTS

[1,2]LASMEDI AFUAN, [3]AHMAD ASHARI, [4]YOHANES SUYANTO
[1] Department of Informatics, Universitas Jenderal Soedirman, Central Java, Indonesia

[2,3,4] Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta Indonesia

E-mail :[1]lasmedi.afuan@unsoed.ac.id, [3]ashari@ugm.ac.id, [4]yanto@ugm.ac.id

ABSTRACT

The number of documents on the Internet has increased exponentially. Every day, users upload various documents to the Internet. This raises a problem, how to find content documents that are relevant to user queries. Information Retrieval (IR) become a useful thing to retrieve documents. However, IR still uses a keyword-based approach to content search that has limitations in displaying the meaning of the content. Often, keywords are used mismatch and miss concept with a collection of documents. As a result, IR displays documents that are not relevant to the context of the information needed. To overcome these limitations, this study has applied Ontology-based IR. The dataset used in the study is 100 learning documents in the field of Informatics which include lecture material, practicum modules, lecturer presentations, proceedings articles, and journals. IR performance evaluation is done by comparing ontology-based IR with classical IR (keyword based). We evaluate IR performance by executing ten queries for testing. Documents that retrieves by query execution are calculated for performance by using Precision, Recall, and F-Measure evaluation metrics. Based on IR performance evaluation, obtained average recall, precision and f-measure values for ontology-based IR of 88.11%, 83.38%, and 85.49%. Meanwhile, IR classics obtained average recall, precision, and f -measure 78.70%, 70.96%, and 74.47%. Based on the values of Recall, Precision, and F-Measure, it can be concluded that the use of ontology can improve relevance document.
Keywords: *Information Retrieval, Ontology, Learning Document, Precision, Recall, F-Measure.*

1. INTRODUCTION

The internet is a very large collection of information. Every day, users upload various documents to the internet that cause problems for users, such as how to find information or documents that are relevant to user needs. Users need special techniques to retrieve documents that are relevant to their needs. IR is one technique that can be used. IR is the process of finding data (usually documents) in the form of text in accordance with the information needed from a collection of documents stored on a computer [1].

IR aims to take documents that are relevant to user needs effectively and efficiently. IR searches for unstructured and semi-structured data sets (such as web pages, documents, images, and videos) [2]. Currently, IR uses keywords to search for document content. In keyword-based IR, when a user enters a query into the system, the system matches the keyword with the document collection content. Often, keyword-based IR provides search results that are not relevant to the context of the information needed. This is caused by a search process that does not consider the context of information, but only matches the words in the document with the keywords entered. As a result, IR displays information that is less relevant to user needs.

To overcome these limitations, in this study we applied an ontology-based IR. Ontology provides shared knowledge about a particular domain and can be reusable. The research question is whether the use of ontology in IR can increase the value of precision and recall? And our hypothesis is using ontology can solve IR classic problems and it can improve precision and recall values. The main contribution of this research is a model of ontology-based information retrieval, which uses to overcome the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

main issues in classical IR. Ontology-based IR can use for search document not only by keyword, but it can search document by the context or the meaning.

This paper is organized as follows. In section 2 discuss some related work. In section 3, present the proposed methodology. In section 4, discuss the results obtained from the evaluation approach. Finally, section 5 concludes.

2. RELATED WORK

Research on IR has been performed by several researchers, based on the studies that have been carried out there are several classical IR models that have been proposed i.e. Boolean Model, Vector Space Model, and Probabilistic Model [1]. At present, on IR there are two important problems, i.e. how to display retrieving relevant documents and how to index documents [2]. IR research has been carried out at different levels to improve the relevance of documents retrieval, including [3] adapt the classic VSM model for ontology-based information retrieval. In this research, the retrieval stage adapts the classic vector space model, including annotation weighting algorithm, and a ranking algorithm. Researches by [4][5][6][6][7][8] proposed query expansion to improve IR performance.

In addition, research conducted by [9] uses a semantic approach to extracting information from the web. Research conducted by [10] proposes HMM (Hidden Markov Model) for IR in polyphonic music. From the results of the research shows that the HMM algorithm works well on a complete music database, but less for small databases. The author [1] proposes an aggregation operator with Modeling as MCDM problem.

Based on the review literature that has been done, the IR research that has been conducted generally uses keywords (keyword based) in conducting content searches, thus creating limitations in displaying and exploring the conceptualization and meaning of content needed by users. Therefore, this study applies the use of ontology to IR. Ontology is used at the indexing stage.

3. PROPOSED WORK

In this study, we propose an ontology-based IR. In general, the architecture of IR-based ontology is shown in Figure 1. In out In the proposed architecture there is five main processes, i.e. document preprocessing, indexing, querying, searching, and ranking.

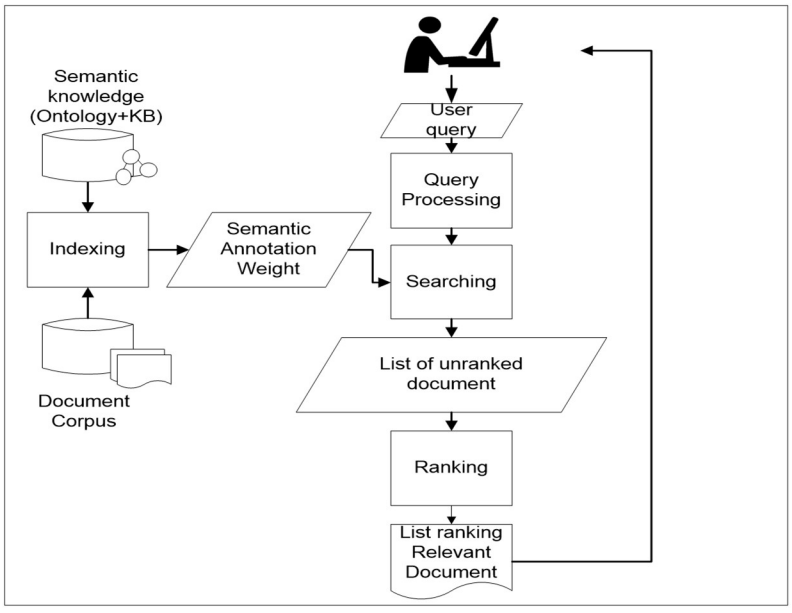

Figure 1: The architecture of the proposed model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Ontology Design

We design ontologies using the approach proposed by [11] and the steps we have followed are:

Step 1: the domain definition and the domain scope.

- The covered domain by our ontology is learning document in computer science or informatics
- The ontology will be used by the lecturer, student, and domain researchers via the search engine.
- The ontology maintenance will be ensured by the specified domain experts.

Step 2: considering the possibility of reusing the existing ontologies.

We perform survey about ontology in computer science or informatics domain.

Step 3: enumerate the most important terms of Ontology.

Due to the high number of terms to be treated in our ontology, we cannot mention them all in this paper. Particular terms in our ontology are shown in Figure 2.

Step 4: define classes and hierarchy

We define class and hierarchy of our ontology, there are several class and subclass that we define.

Step 5 and 6: define the classes properties

We define object properties and data properties for out ontology.

Step 7: Creating the instances

Our ontology concepts represent terms related to the computer science and informatics domain and there no instance for these terms. For example the concept "Database" has no instance.

Furthermore, Classes, subclasses, Object properties, Data properties, and instances that have been generated will be implemented using Protégé tools. This choice is supported by several reasons:

- Protégé is free and open source editor
- It can use for defining a class of concepts.
- Ontologies can be edited interactively within Protégé.
- Ability to archive ontologies and knowledge bases in a variety formats.

The ontology construction results are shown in Figure 2.

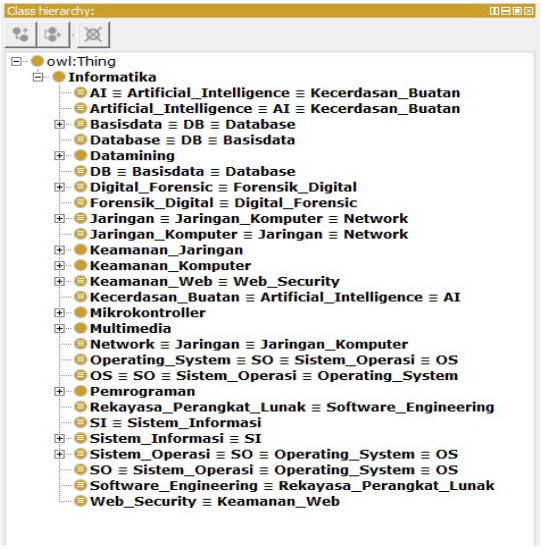

Figure 2: Ontology design.

3.2 Document Preprocessing

Document preprocessing process is carried out, the results of this process will be used at the indexing process. The preprocessing document process steps are shown in Figure 3. The document preprocessing process constitutes several steps:

1. Extracting documents, documents used in the format (.pdf), so that processing can be done, the document is extracted and saved as a document (.txt) format. The software used by PDFBox, we do embed PDFBox into python.
2. Lowercase, a document that has been extracted into a text (.txt) format, the next step is to change all letter characters to lowercase.
3. Stopword removal, removes vocabulary that is not a feature (unique word) of a document.
4. POS tagging, this process give word class labels, including verbs, nouns, and others. This study uses word classes based on the research proposed by [12].

The results of preprocessing are a collection of terms or phrases which are then stored in a relational database and also stored in a text (.txt) format document that will be used at the indexing process.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

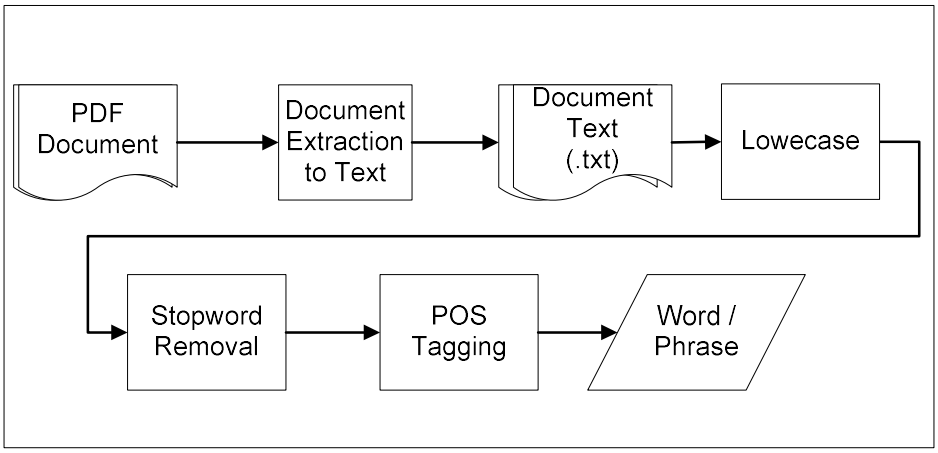

Figure 3: Document Preprocessing Process.

3.3 Indexing Process

Indexing is a very important step in IR, indexing represents a document. Indexing functions to build indexes of documents. the index will be used during the searching phase. In this study, indexing is done using semantic annotations. The main process of semantic annotation is mapping between the term/phrase in the document with semantic ontology entities (knowledge base). The ontology that has been made as shown in Figure 4, Each semantic entity on ontology is given more than one label. The label functions as a textual representation of each semantic entity. The overall semantic annotation process is shown in Figure. 4, and consists of the following steps for every semantic entity in every ontology:

1. Load the information of a semantic entity, that is, extracting the textual representation of the selected entity. Each entity may have one or more textual representations in the ontology. SPARQL perform extraction using *rdfs:Label* vocabulary. For example, an operating system entity has more than one textual representation, i.e. the operating system and OS. The results of textual representation extraction are stored in semantic entity tables in the database.
2. Perform semantic annotation, Semantic annotations are carried out by mapping between the appearance of textual representations with terms/phrases that are in the document preprocessing. In this process, we use the Whoosh program with the results of textual representation as search keywords in the document to be searched. The results of the semantic annotations are shown in Table 1.

Table 1: Semantic annotation results.

No	Semantic entities/concept	Document
1	E1	DOC01, DOC33, DOC93, DOC03
2	E2	DOC30, DOC90
3	E4	DOC09, DOC69, DOC99, DOC50, DOC43
4	E5	DOC79, DOC19, DOC10, DOC100, DOC70
5	E7	DOC07, DOC97, DOC41, DOC50
..	..	..
1463	E142	DOC15, DOC75, DOC10, DOC100, DOC70, DOC11, DOC71, DOC13, DOC73

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

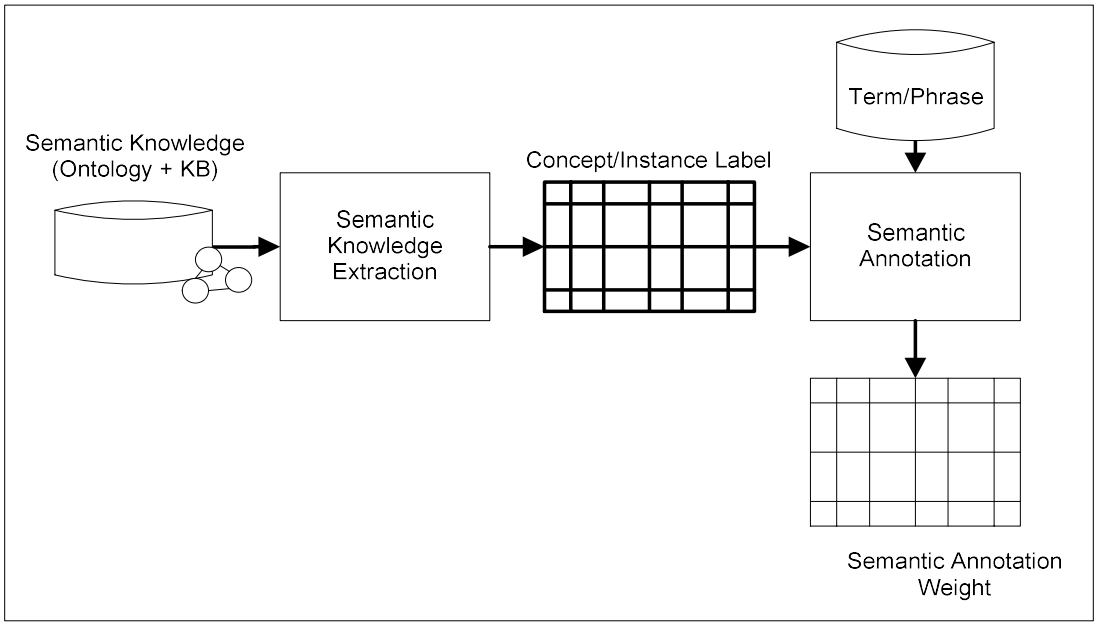

Figure 4: Semantic annotation process.

3. Furthermore, We perform weighting semantic annotations. For weighting semantic annotations is done by modifying the TF-IDF to define the calculation of weights of instances in a document using equations (1)

$$d_x = \frac{freq_{x,d}}{max_y freq_{x,d}} .\log \frac{|D|}{n_x} \quad (1)$$

Where $freq_{x,d}$ is the number of occurrences in d of the keywords attached to x, $max_y freq_{y,d}$ is the frequency of the most repeated instance in d, n_x is the number of documents annotated with x, and D is the set of all documents in the search space. The semantic annotation results shown in Table 2.

Table 2: Semantic annotation results.

No	Semantic entities	Document	Weight	No	Semantic entities	Document	Weight
1	E1	DOC01	0.321	12	E5	DOC79	0
2	E1	DOC33	0.458	13	E5	DOC19	0
3	E1	DOC93	0.076	14	E5	DOC10	0
4	E1	DOC03	0.076	15	E5	DOC100	0
5	E2	DOC30	0.163	16	E5	DOC70	0
6	E2	DOC90	0.163	17	E7	DOC07	0.279
7	E4	DOC09	0.975	18	E7	DOC97	0.611
8	E4	DOC69	1.062	19	E7	DOC41	0.401
9	E4	DOC99	0.975	20	E7	DOC50	0.169
10	E4	DOC50	0.079	..	..	..	..
11	E4	DOC43	0.041	1463	E142	DOC73	0

3.4 Querying Process

Users enter queries into the system, and then the system preprocesses the query. Next, the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system matches the query results of preprocessing with the concept or entity that has been stored in the annotation table.

3.5 Searching And Ranking Process

The searching process performed by mapping between semantic entities or concepts extracted from user queries with the semantic index results from the indexing process.

The ranking process performed using the Vector Space Model (VSM) and TF-IDF based on cosine similarity. Ranking performs on documents produced at the searching process. Each document is represented as a vector, where elements of each vector are weights of entity semantic annotations from documents [3]. Queries are also represented as vectors, where elements of vectors are semantic weights of entities that are related to query variables [13]. Researchers [3] define the size of similarity between a document d and query q as cosine similarity with the equation (2)

$$sim(d,q)=\frac{\vec{d}.\vec{q}}{|\vec{d}|.|\vec{q}|.} \quad (2)$$

4. EVALUATION

4.1 Dataset

This study uses 100 Indonesian documents covering lecture materials, practical modules, lecturer presentations, article proceedings, and journals. The document is specific to the domain in the field of Informatics. The documents are grouped into twelve categories of material as presented in Table 3.

Table 3: Documents category.

No	Categories Name	Document Count
1	Database	13
2	Information Technology	16
3	Computer Network	23
4	Information Systems	20
5	E-Commerce	2
6	Operating System	3
7	Object-Oriented Programming	11
8	Computer Programming	8
9	Computer Security	1
10	Artificial Intelligence	1
11	Text Mining	1
12	Ontology	1
	Total of documents	100

4.2 Evaluation Method

Ontology-based IR evaluation, we prepared ten Indonesian queries as examples shown in Table 4. We placed the appropriate keyword queries for use in the evaluation. Then, calculate the number of correct documents to be retrieved, for each query. Finally, we run the query and calculate the performance using Precision, Recall, and F-Measure evaluation metrics. Evaluation using the value of recall and precision done to determine the level of relevance and accuracy of the system in searching for information requested by the user. In the evaluation of the relevance level, the recall (R) value represents the value that shows the rate of returns returned by a system. This value is obtained by comparing the number of relevant items returned by the system with the total number of relevant items present in the system collection as in Equation (3). A good system not only showed by greater recall value. The highest recall value is 1, which means that all documents in the collection are found.

Recall

$$R=\frac{TP}{TP+FN} \quad (3)$$

Precision value (P) shows the level of accuracy of a system to return relevant information to the user. This value is obtained by comparing the number of relevant items returned with the total number of items returned as in Eq. (4). The greater the precision value of a system, the system can be said to be good. The highest precision value is 1, which means all documents found are relevant.

Precision

$$P=\frac{TP}{TP+FP} \quad (4)$$

F-Measure is a combination of recall and precision that takes the weight of the harmonic mean. F-Measure value will be high if recall and precision have high value, to calculate F-Measure used equation (5)

F-Measure

$$F\text{-}Measure=2.\frac{PR}{P+R} \quad (5)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: IR Evaluation queries

Queries	Label/caption
Q1	DATABASE
Q2	MYSQL
Q3	NETWORK
Q4	DML
Q5	WEBSITE
Q6	HTML
Q7	JAVA
Q8	PROTOCOL
Q9	TOPOLOGY
Q10	PHP

4.3 Testing Results

The testing results using ten queries entered into the system, then calculating the precision, recall, and f-measure. The results of the testing are shown in Table 5. In table 5, we compare the IR Classics that have been done in the prior work with the Ontology-based IR that we are doing.

Table 5: The results of testing.

Queries	Ontology-based IR				Classical IR			
	Document retrieved by system	TP	FN	FP	Document retrieved by system	TP	FN	FP
Q1	9	8	1	1	9	8	1	1
Q2	10	7	1	3	10	5	4	5
Q3	10	10	1	1	57	50	5	7
Q4	6	6	1	1	6	5	1	1
Q5	8	6	1	2	8	5	1	3
Q6	7	6	1	1	7	5	1	2
Q7	10	10	1	1	23	20	2	3
Q8	10	9	1	1	15	9	3	6
Q9	10	9	1	1	16	11	3	5
Q10	9	6	1	3	8	4	3	4

From Table 4, precision, recall, and f-measure calculations are perform using Equations (4), (5), and (6). The calculation results are shown in Table 6.

Table 6: The calculation results (Recall, Precision, F-measure)

Queries	Ontology-based IR			IR Classic		
	Recall (%)	Precision (%)	F-Measure	Recall (%)	Precision (%)	F-Measure
Q1	88.89	88.89	88.89	88.89	88.89	88.89
Q2	87.50	70.00	77.78	55.56	50.00	52.63
Q3	90.91	90.91	90.91	90.91	87.72	89.29
Q4	85.71	85.71	85.71	83.33	83.33	83.33
Q5	85.71	75.00	80.00	83.33	62.50	71.43
Q6	85.71	85.71	85.71	83.33	71.43	76.92
Q7	90.91	90.91	90.91	90.91	86.96	88.89
Q8	90.00	90.00	90.00	75.00	60.00	66.67
Q9	90.00	90.00	90.00	78.57	68.75	73.33
Q10	85.71	66.67	75.00	57.14	50.00	53.33
Average	88.11	83.38	85.49	78.70	70.96	74.47

From Table 6 we describe a comparison chart for recall, precision, and f-measure between Ontology-based IR and Classical IR. It is shown in Figure 5.6, and 7.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

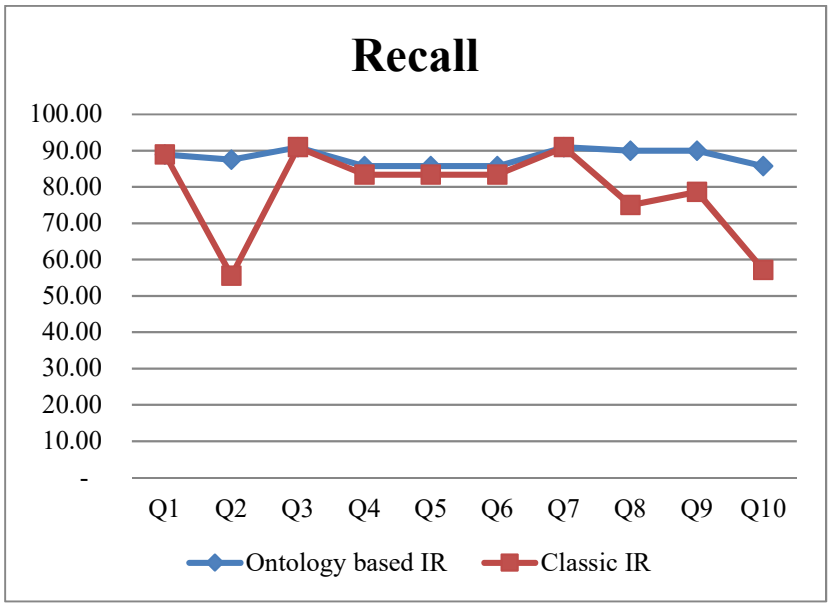

Figure 5: Recall.

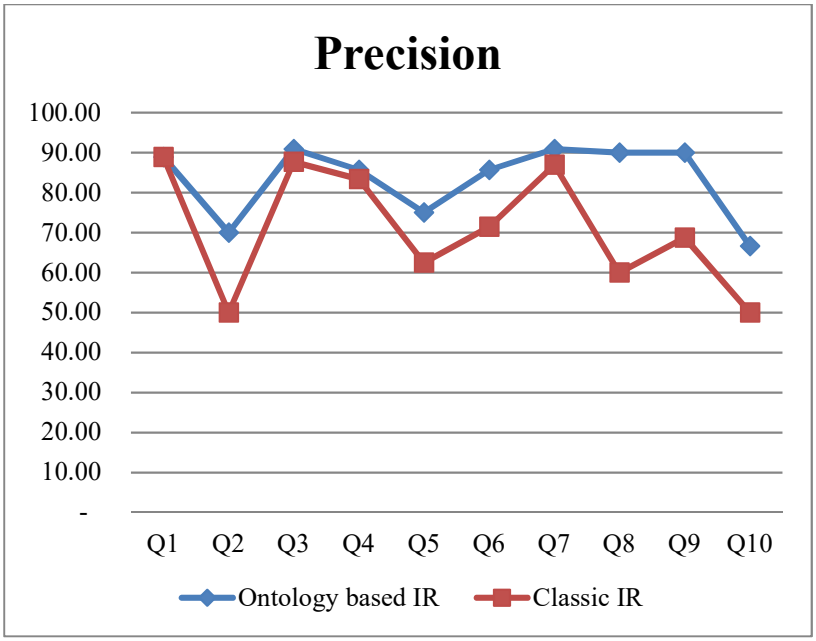

Figure 6: Precision.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

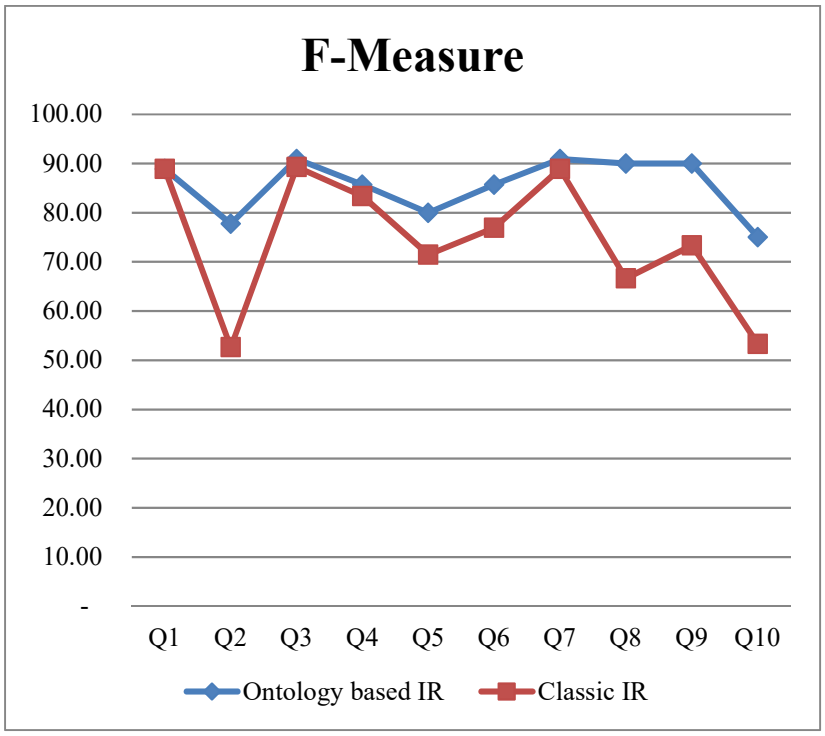

Figure 7: F-Measure.

5. CONCLUSION

In this study, we have implemented an ontology-based IR for learning documents in the informatics domain. Based on IR performance evaluation, obtained average recall, precision, and f-measure values for ontology-based IR of 88.11%, 83.38%, and 85.49%, for Classical IR obtained average recall, precision, and f-measure 78.70%, 70.96%, and 74.47%. Based on the values of Recall, Precision, and F-Measure, it can be concluded that the use of ontology on IR can improve the relevance of documents.

For future work, we will use query expansion with integration between association rules and ontology for improving relevance document in IR.

6. ACKNOWLEDGMENTS

Authors would like to thank the reviewers for detailed, accurate and helpful comments. Our thanks to the Ministry of Research, Technology and Higher Education for financial support, so we can complete this research.

REFERENCES

[1] S. Marrara, G. Pasi, and M. Viviani, "Aggregation operators in Information Retrieval," *Fuzzy Sets Syst.*, vol. 1, pp. 1–17, 2017.

[2] B. M. Sanderson and W. B. Croft, "The History of Information Retrieval Research," in *IEEE*, 2012, vol. 100, pp. 1444–1451.

[3] P. Castells, M. Fernandez, D. Vallet, M. Fernández, and D. Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 2, pp. 261–272, 2007.

[4] A. Gomathi, J. Jayapriya, G. Nishanthi, K. S. Pranav, and P. K. G, "Ontology Based Semantic Information Retrieval Using Particle Swarm Optimization," *Int. J. Appl. Inf. Commun. Eng.*, vol. 1, no. 4, pp. 5–8, 2015.

[5] D. Zhou, S. Lawless, J. Liu, S. Zhang, and Y. Xu, "Query Expansion for Personalized Cross-Language Information Retrieval," *Int. Work. Semant. Soc. Media Adapt. Pers.*, 2015.

[6] M. Amina, L. Chiraz, and Y. Slimani, "Short Query Expansion for Microblog Retrieval," *Procedia - Procedia Comput.*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Sci., vol. 96, pp. 225–234, 2016.

[7] M. C. Di. Galiano, M. . M. Valvidia, and L. . U. Lopez, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Comput. Biol. Med.*, vol. 39, pp. 396–403, 2009.

[8] L. C. Chen, W. T. Chao, and C. J. Hsieh, "A Novel Query Expansion Method for Military News Retrieval Service," *IEEE*, pp. 183–186, 2014.

[9] F. Dahak, M. Boughanem, and A. Balla, "A probabilistic model to exploit user expectations in XML information retrieval," *Inf. Process. Manag.*, vol. 53, no. 1, pp. 87–105, 2017.

[10] S. Chithra, M. S. Sinith, and A. Gayathri, "Music Information Retrieval for Polyphonic Signals using Hidden Markov Model," *Procedia - Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 381–387, 2015.

[11] N. F. Noy and D. L. Mcguinness, "Ontology Development 101 : A Guide to Creating Your First Ontology," pp. 1–25, 2000.

[12] A. F. Wicaksono and P. Ayu, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," *Prooceedings 4th Int. MALINDO (Malay Indones. Lang. Work.*, no. January 2010, pp. 1–7, 2010.

[13] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced Information Retrieval: An ontology-based approach," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 9, no. 4, pp. 434–452, 2011.