Using of Jaccard Coefficient for Keywords Similarity

Suphakit Niwattanakul*, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu

***Abstract*—Presently, information retrieval can be accomplished simply and rapidly with the use of search engines. This allows users to specify the search criteria as well as specific keywords to obtain the required results. Additionally, an index of search engines has to be updated on most recent information as it is constantly changed over time. Particularly, information retrieval results as documents are typically too extensive, which affect on accessibility of the required results for searchers. Consequently, a similarity measurement between keywords and index terms is essentially performed to facilitate searchers in accessing the required results promptly. Thus, this paper proposed the similarity measurement method between words by deploying Jaccard Coefficient. Technically, we developed a measure of similarity Jaccard with Prolog programming language to compare similarity between sets of data. Furthermore, the performance of this proposed similarity measurement method was accomplished by employing precision, recall, and F-measure. Precisely, the test results demonstrated the awareness of advantage and disadvantages of the measurement which were adapted and applied to a search for meaning by using Jaccard similarity coefficient.**

***Index Terms*—Keyword, Similarity, Jaccard Coefficient, Prolog programming language**

I. Introduction

MODERN information retrieval can be accessed from services of Search Engines such as Google, Yahoo, Bing, and AltaVista. The users can search for information in multimedia formats such as text, audio, still images, and moving images [1] by looking up for keywords appeared in any documents and/or files stored in different formats, such as HTML, MS Word, MS Excel, PDF, and images. These documents and/or files, which are distributed over a large data source, will be stored on the Internet. As a result in wide range information searching, searchers are not able to access the whole site causing incapability to obtain specific information. Additionally, searching results of meaning similarity and relation to keywords in some cases might not display required documents that do not contain specific keywords inputted. Thus, they are not able to find the document or web page they need [2]. This can be a result in lacking of searching technique or knowledge of how to use a specific keyword or keywords and search process.

Keyword search is the simplest form of the most popular query method for search engine in information systems [1]. It contains a single keyword or multiple keywords and a sort phrase. In a single keyword search, a particular word in the document will be displayed such as in a case of searching for sugar-producing crops. Keywords are specific words that can be sugar cane or we can query with the keyword in other forms to allow users to easily find the needed information quickly. The first significant issue that needs to consider is the technique used to measure the similarity between a user-specified key and the index finger to indicate directly to the required information.

From the study of [3], they researched on search engine optimization services by analyzing manifest page display names with proximity comparison between user's request and each document represented in a database format. The document, that is most similar to the request, is query answers. General information retrieval systems use principle of words frequency that appears in documents with the weight of a variable in the specified document and the proximity of user's request. Nevertheless, page name search in the study cannot apply the abovementioned variables because the frequency of words in the document is analyzed and displayed as prominent name only. Thus, the frequency of the variable cannot be used to specify the proximity of the data. In this paper, they used Jaccard similarity coefficient method as it is popularly used to compare the proximity of the data in the process data (Data Clustering) [4]. This method can be given the proximity of the two data sets efficiently without the use of data redundancy. The results showed that when prominent document names were analyzed, the represented documents were displayed correctly. This results in a higher precision of the system and the smaller database than a typical search page with other services. In [5] said that the search process commences from importing users' queries to compare with the database. In case of input keyword matches with the index of words in the database, those words can be accounted for the main keywords displayed in that search process. Nonetheless, if a query does not match any index in the database, the process of similarity measurement can be proceeded to scrutinize the most similarity of the words stored structurally in the database such as Keywords, Similar Words, Broader Term (BT), Narrower Term (NT) and Related Term (RT), by using Jaccard similarity coefficient as displayed below.

Manuscript received December 8, 2012; revised January 10, 2013.

Suphakit Niwattanakul,Ph.D. is an instructor with the School of Information Technology, Suranaree University of Technology, Thailand (email: suphakit@sut.ac.th).

Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu are a doctoral student with the School of Information Technology, Suranaree University of Technology, Thailand.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$Jaccard\ sim(A,B) = P(A \cap B) | P(A \cup B) \quad (1)$$

A search keyword can be used effectively when similarity is computed within acceptance criteria which are equal to 0.75. In [6] describes a document retrieval system as the Information Retrieval (IR) System which is designed to retrieve documents by a user query from large archive documents. The system is primarily responsible to document operations, creates a document representation or an index, query operations and representation, and searches documents by comparing the similarities (Similarity Computation) of a keyword and the document agents. Results of the system are a list of documents sorted (Ranking) by the similarity of documents displayed to users. Therefore, this research paper focused on measuring the similarity of the keyword using Jaccard Coefficient that was developed to measure the similarity of the Jaccard with Prolog programming language as a linear function. The test result was to determine the advantage and disadvantages of Jaccard similarity coefficient method that can be adapted and applied to the search for semantic data access and retrieval.

II. METHODOLOGY

A. The data relationship between the information.

This research paper was classified into two parts: 1) the information prepared as words (Here I use the word "words" to mean the set of words or phrases) which were grammatically correct. The keywords were taken from the thesaurus of agricultural Thailand in the farm topic section of 100 words; and 2) the information that was not grammatically correct was tested in three groups (the misspelled words, crashed words, and over-typed words) by users. These words were also determined by the researchers. The example of words is displayed in Fig. 1 below.

Keywords Search	Index word
Correct grammar words	อ้อย
อ้อย มันสำปะหลัง ข้าวโพด สับปะรด	มันสำปะหลัง
ถั่วเหลือง ถั่วเขียว ถั่วลิสง	ข้าวโพด
	สับปะรด
Misspelled words	ถั่วเหลือง
อ้นย มันสำปังหลัง ข้ววโพด	ถั่วเขียว
สับปะนด ถั่วเกลือง ถั่วเขือว	ถั่วลิสง
	น้ำตาล
Crashed words	ข้าว
อ้ย มันสำปหลัง ข้าโพด	ข้าวเหนียว
สับปะร ถั่วเหลอง ถั่วขียว	คาร์โบไฮเดรต
	แป้ง
Over-typed words	พันธุ์ข้าว
อ้อยย มันสำปปะหลัง ข้าววโพด	ข้าวหอมมะลิ
สสับปะรด ถั่วเหลืองง ถั่ววเขียว	รวงข้าว

Fig. 1. Sample words that appeared in the index and the query (correct grammar words, misspelled words, crashed words, and over-typed words).

B. A measure of similarity of the search words.

1. The determination of the association between two words with Jaccard coefficient.

Jaccard index is a name often used for comparing similarity, dissimilarity, and distance of the data set.

Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all divided by the number of properties as shown below.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Jaccard distance is non-similar measurement between data sets. It can be determined by the inverse of the Jaccard coefficient which is obtained by removing the Jaccard similarity from (1). It is equal to a number of features that are all minus by number of features that are common to all divided by the number of features as presented below.

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (3)$$

This is the similarity of asymmetric binary attributes. Viewing the properties of an object in a binary format enables user to measure the similarity more easily by determining the Objects A and B comprising “n” features. The Jaccard similarity uses a measure of the share properties of both Objects A and B whereas all of the Objects A and B given by 0 and 1 respectively.

2. The calculation of search words to identify similarity.

To illustrate more clearly, the following example displayed in a form of set diagrams known as Venn Diagrams which were determined as Set A for "มันสำปะหลัง" and Set B for "มันฝรั่ง" as shown in Fig. 2.

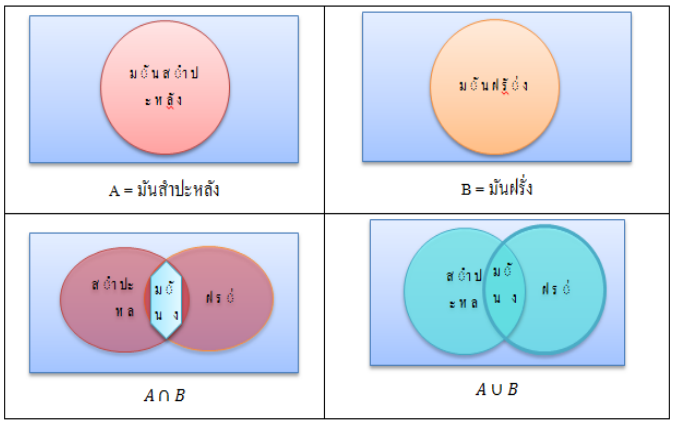

Fig. 2. Set Diagrams of the calculation of Jaccard similarity coefficient.

From the above illustration, it can be used to calculate the Jaccard similarity coefficient as presented below.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{น, ้, น, ้\}|}{|\{ส, ้า, ป, ะ, ร, ด, น, ้, ป, ะ, ร, ้\}|} = \frac{4}{13} = 0.30769 \quad (4)$$

C. Performance Evaluation.

We evaluated the similarity performance of search words by using the precision, recall, and F-measure. It was calculated by the following.

Precision

P = (Number of accurate results*100)/Total of answers retrieving by the system

Recall

R = (Number of accurate results*100)/Total of accurate results from raw data

F-measure

F =(2 x Precision x Recall)/(Precision+Recall)

III. RESULT AND DISCUSSION

A. Coding of Coefficient test program for Jaccard Similarity with Prolog programming language.

On the one side, Prolog programming language as Inference Engine is a program with a capability to learn whatever commands inputted by the developers. On the other side, it is the language of artificial intelligence. The grammar can be learned relatively in a short period of time. It is ideal for developing logical solutions, artificial intelligence, and computational linguistics. Sample codes are shown in Fig. 3 and Fig. 4.

```
jaccardsCoefficient(T1,T2,R) :- name(T1, CharList1),
                                name(T2, CharList2),
                                intersection(CharList1,CharList2,RL),
                                sizeList(RL,IR),
                                union(CharList1,CharList2,RL2),
                                sizeList(RL2,UR),
                                R is (IR)/(UR).

union([],L,R)   :- union(L,[],R).
union([H|TL],L,R)   :- clearDuplicate(H,TL,NTL),
                       clearDuplicate(H,L,NL),
                       union(NTL,NL,RL),
                       R = [H|RL].

intersection([],_,[]).
intersection([H|TL],CharList2,R)    :- (   inList(H,CharList2)
                                       -> clearDuplicate(H,CharList2,NewCharList2),
                                          intersection(TL,NewCharList2,TR),
                                          R = [H|TR]
                                       ;  NewCharList2 = CharList2,
                                          intersection(TL,NewCharList2,TR),
                                          R = TR  ).

clearDuplicate(_,[],[]).
clearDuplicate(C,[H|TL],R)  :- clearDuplicate(C,TL,TR),
                               (   C = H
                                  -> R = TR
                                  ;  R = [H|TR]  ).

sizeList([],0).
sizeList([H|TL],R)  :- clearDuplicate(H,TL,TR),
                       sizeList(TR,RI),
                       R is RI + 1.
```

Fig. 3. Sample codes of Jaccard Similarity Coefficient in Prolog programming language

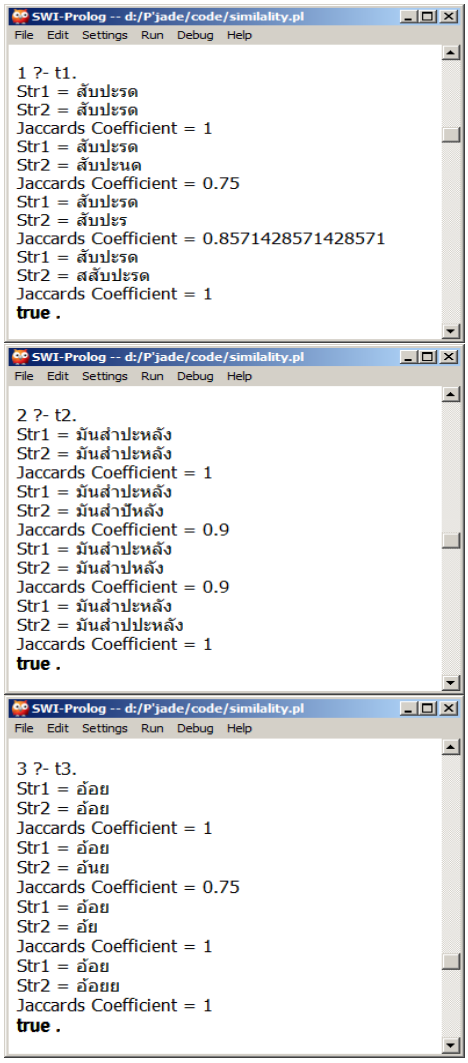

Fig. 4. Sample reports of execution logs.

B. Comparison of the normal test.

Table I presented the result of the accuracy testing of Jaccard similarity coefficient on the data sets with correct grammar syntax.

TABLE I
JACCARD SIMILARITY COEFFICIENT WITH THE CORRECT GRAMMAR SYNTAX.

	อ้อย	มันสำปะหลัง	ข้าวโพด	สับปะรด	ถั่วเหลือง	ถั่วเขียว	ถั่วลิสง
อ้อย	1.000						
มันสำปะหลัง	0.000	1.000					
ข้าวโพด	0.111	0.000	1.000				
สับปะรด	0.000	0.308	0.077	1.000			
ถั่วเหลือง	0.083	0.250	0.063	0.063	1.000		
ถั่วเขียว	0.100	0.059	0.154	0.071	0.385	1.000	
ถั่วลิสง	0.000	0.286	0.071	0.154	0.500	0.333	1.000

C. A comparative test of the error.

Tables II to VI illustrated testing results of the accuracy of Jaccard similarity coefficient with corrected words, misspelled words, crashed words, and over-typed words.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II
JACCARD SIMILARITY COEFFICIENT WITH AN ERROR.

	Correct word อ้อย	Misspelled word อ้นย	Crashed word อ้ย	Over-typed word อ้อยย
อ้อย	1.000	0.750	1.000	1.000
มันสำปะหลัง	0.000	0.077	0.000	0.000
ข้าวโพด	0.111	0.100	0.111	0.111
สับปะรด	0.000	0.000	0.000	0.000
ถั่วเหลือง	0.083	0.077	0.083	0.083
ถั่วเขียว	0.100	0.091	0.100	0.100
ถั่วลิสง	0.000	0.000	0.000	0.000

TABLE III
JACCARD SIMILARITY COEFFICIENT WITH AN ERROR.

	Correct word มันสำปะหลัง	Misspelled word มันสำป์หลัง	Crashed word มันสำปหลัง	Over-typed word มันสำปปะหลัง
อ้อย	0.000	0.000	0.000	0.000
มันสำปะหลัง	1.000	0.900	0.900	1.000
ข้าวโพด	0.000	0.000	0.000	0.000
สับปะรด	0.308	0.231	0.231	0.308
ถั่วเหลือง	0.250	0.267	0.267	0.250
ถั่วเขียว	0.059	0.063	0.063	0.059
ถั่วลิสง	0.286	0.308	0.308	0.286

TABLE IV
JACCARD SIMILARITY COEFFICIENT WITH AN ERROR.

	Correct word ข้าวโพด	Misspelled word ข้ววโพด	Crashed word ข้าโพด	Over-typed word ข้ววโพด
อ้อย	0.111	0.125	0.125	0.111
มันสำปะหลัง	0.000	0.000	0.000	0.000
ข้าวโพด	1.000	0.857	0.857	1.000
สับปะรด	0.077	0.083	0.083	0.077
ถั่วเหลือง	0.063	0.067	0.000	0.063
ถั่วเขียว	0.154	0.167	0.077	0.154
ถั่วลิสง	0.071	0.077	0.000	0.071

TABLE V
JACCARD SIMILARITY COEFFICIENT WITH AN ERROR.

	Correct word สับปะรด	Misspelled word สับปะนด	Crashed word สับปะร	Over-typed word สสับปะรด
อ้อย	0.000	0.000	0.000	0.000
มันสำปะหลัง	0.308	0.417	0.333	0.308
ข้าวโพด	0.077	0.077	0.000	0.077
สับปะรด	1.000	0.750	0.857	1.000
ถั่วเหลือง	0.063	0.063	0.067	0.063
ถั่วเขียว	0.071	0.071	0.077	0.071
ถั่วลิสง	0.154	0.154	0.167	0.154

TABLE VI
JACCARD SIMILARITY COEFFICIENT WITH AN ERROR.

	Correct word ถั่วเหลือง	Misspelled word ถั่วเกลือง	Crashed word ถั่วเหลอง	Over-typed word ถั่วเหลืองง
อ้อย	0.083	0.083	0.091	0.083
มันสำปะหลัง	0.250	0.176	0.267	0.250
ข้าวโพด	0.063	0.063	0.067	0.063
สับปะรด	0.063	0.063	0.067	0.063
ถั่วเหลือง	1.000	0.818	0.900	1.000
ถั่วเขียว	0.385	0.385	0.417	0.385
ถั่วลิสง	0.500	0.500	0.545	0.500

The results showed that Jaccard similarity coefficient were in between 0 and 1. A value of 0 indicates that there is no similarity whereas a value of 1 indicates a similarity.

Table VII displayed the analysis of the similarity coefficient by deploying precision, recall, and F-measure for the performance measurement. If a value greater than 0.55 means that the word is selected and the rest can be interpreted as not selected.

TABLE VII
THE SIMILARITY COEFFICIENT GREATER THAN 0.55

Keyword	Precision	Recall	F-Measure
Crashed words	93.75	100	96.77
Over-typed words	85.71	100	92.31
Misspelled words	93.75	100	96.77

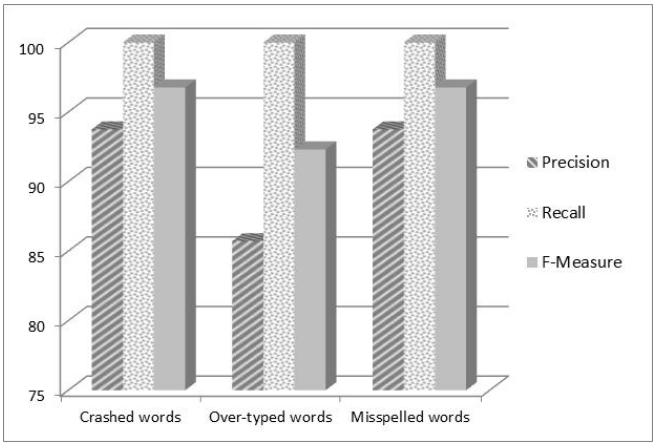

Fig. 5. The results with similar values greater than 0.55

Fig. 5 indicates that the performance can be estimated very accurate and stable at a high performance in all cases.

Dissimilarity, Table VIII presented the similarity coefficient results of the performance measurement using Precision, Recall, and F-measure. If a value of similarity is greater than 0.55 and not equal to 1.0, it means that the word is selected and the rest can be interpreted as not selected. Moreover, keywords or words that is not selected and gives the similarity value of 1.0, it referred as incorrect keywords or words.

Additionally, Fig. 6 illustrates that Jaccard similarity coefficient had error values when there was an event of typing the same word repeatedly which caused the result remained in the highest value or 1.0. This means that the algorithm of Jaccard similarity coefficient cannot verify the existence of duplicate samples such as "อ้อย" "อ้ยย", "อ้ย", and "ออ้อย" which were all equals to 1.0.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE VIII
THE SIMILARITY COEFFICIENTS GREATER THAN 0.55 AND LESS THAN 1.0

Keyword	Precision	Recall	F-Measure
Crashed words	93.33	93.33	93.33
Over-typed words	66.67	33.33	44.44
Misspelled words	93.55	96.67	95.08

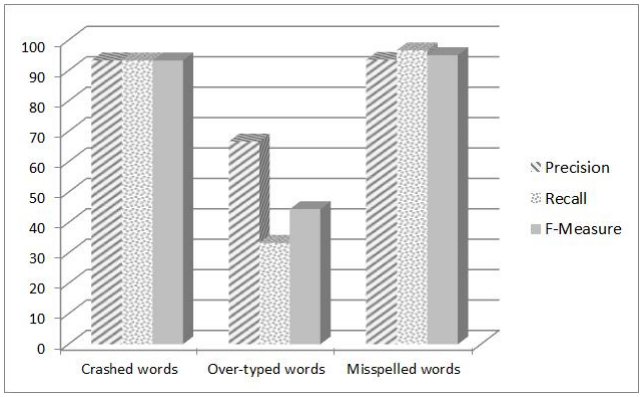

Fig. 6. The results with similarity values greater than 0.55 and less than 1.0

In case of over-typed words as shown in Fig. 6, the prediction accuracy was declined and the stability was obviously dropped. This indicated that over-typed words were neglected in the measurement of the similarity with Jaccard coefficient.

IV. CONCLUSIONS AND FUTURE RESEARCH

This research paper tested the algorithm to find about Jaccard similarity coefficient by measuring the similarity in the correct grammar syntax and the test of similarity in terms of an error by developing the tests with Prolog programming language. The results showed that the test method by Jaccard coefficient can perform well in measuring the similarity of words when comparing with each letter of the word. Particularly, each letter can switch positions and counted as the same words. Nevertheless, this method is not able to detect the over-type words in the data sets. In conclusion, Jaccard similarity coefficient is suitable sufficiently to be employed in the word similarity measurement. In efficiency measurement, the program performance can deal appropriately with high stability when failure and mistake spelling occurred.

The test results also showed some weaknesses of the Jaccard similarity coefficient when measuring similarity of certain words. Therefore, the other algorithms such as Vector Space, Cosine Coefficients, and Engram should be also considered and tested to apply and modify the advantages of each algorithm for semantic search performance and satisfy the need of users.

REFERENCES

[1] Supachai Tungwongsarn. (2010). *System for storage and retrieval of information by computer*. Bangkok: Pithak Printing.

[2] Manusnanth Panyamee, and Somjit Arj-in. (2009). Document clustering results on the semantic web search. In *Proceedings of The 5th National Conference on Computing and Information Technology*. (Page 1). Bangkok: King Mongkut's University of Technology.

[3] Jirus Malawong, Arnonth Roonsawang.(2003). Performance Optimization of name page search service with an analysis of the dominant subject name. In *Proceedings of The 7th National Computer Science and Engineering Conference Named page, Frequent itemset.* Bangkok: Kasetsart University.

[4] Guha.S, Rastogi.R, and Shim.K. (1999). ROCK: A Robust Clustering Algorithm for Categorical Attributes, *in Proceedings of International Conference on Data Engineering (ICDE)*, Sydney, Australia, pp. 512-521.

[5] Suphakit Nivattanakul.(2008). *Access to Knowledge Based-on an Ontology Model.* Ph.D. thesis. University of La Rochelle.

[6] Salton G. (1989). *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)