An IR-Based Semantic Search Approach on Portable Document Format (PDF) Files Using Similarity Measure

AGSUNOD,
John Mark Robert M.

BANTING,
Carl Jayson M.

BRAR,
Harjit S.

CUNANAN,
Patrick Bryan F.

PONAY,
Charmaine S.

Institute of Information and Computing Sciences
University of Santo Tomas, España Blvd., Manila, Philippines
{johnmark.agsunod, carljayson.banting, harjit.brar, patrickbryan.cunanan}.iics@ust.edu.ph,
csponay @ust-ics.mygbiz.com

ABSTRACT

Information Retrieval is finding material, usually documents, of an unstructured nature, usually text, that satisfies an information need from within large collections, usually stored in computer databases.

This research introduces a semantic search engine that is able to generate a list of relevant documents that are related to a certain user query faster than current search engines by implementing the similarity measure in its search process. A familiar user interface would be adopted by the implemented system that is generally known to almost everyone to ensure ease of use, efficiency, and usability.

As for existing semantic search engines, the implemented system would be redesigned in order to perform better in terms of speed and accuracy. This is done by introducing the similarity measure, along with several algorithms into its search process. The evaluation of these factors were calculated using the precision, recall, and F1-measure tools in order to ensure the reliability and efficiency of the implemented system.

The evaluation results show that the accuracy of the implemented system stayed the same and the actual time or speed is decreased by 30%, compared to the existing system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]:
H.3.3 Information Search and Retrieval
I.2 [**Artificial Intelligence**]
I.2.7 Natural Language Processing

General Terms

Algorithms, Semantic search

Keywords

3rd-order Tensor Model, Concept Space, Concept Vector, Concept Window, Similarity Measure, Lesk Algorithm, TF-IDF Algorithm

1. INTRODUCTION

The idea of searching is changing. From the time when search engines were first introduced, a plethora of advancements and additions have been applied to traditional search engines in order to accommodate the rapid influx of new information being made available. In order to cope with this rapid data growth, the concept of information retrieval is developed and deployed. Information retrieval is a technique that mixes computer science and information science. Christopher Manning wrote that Information retrieval (IR) is “finding material, usually documents, of an unstructured nature, usually text, that satisfies an information need from within large collections, usually stored on computers”. Certain applications of Information Retrieval in our activities and in the industry include activities done by paralegals, librarians and other professional searchers (Manning et al., 2009).

As stated by G. Chowdhury, the goal of making an information retrieval (IR) system is to aid in recovering and accessing documents. “IR systems are designed to analyze, process and store sources of information and retrieve those that match the user’s preferred requirements” (Chowdhury, 2010).

The field of information retrieval has been progressively developing along with the development of search engines and the World Wide Web. According to Wall, in the year 1990, the first search engine was brought to life by Alan Emtage, with his creation named Archie. Wall wrote, “It was a pre-Web search engine for locating material in specific public FTP archives. Though it only operated for a short while and updates were ceased later that same year, similar search engines were formed in likeness to Archie, such as Jughead, produced by the University of Minnesota, and Veronica, which was developed by the University of Nevada to search on plain text files” (Wall, 2006). The use of information retrieval has been used in many fields such as medicine and engineering, considering the large amounts of jargons and expansive data gathered and documented on those fields. One area that the research team aims to improve knowledge on is data specific to technology, specifically in the categories of Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search.

With technology specific to Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search as the research domain, the researchers can implement an approach to develop a search engine agent that generates a set of selected documents from a database, which is then processed with IR algorithms to produce a collected list of documents related to the terms determined in the user query.

In the study, where most of our study is based upon, entitled "*A semantic search technique with Wikipedia-based text representation model*" (Hong et al., 2016), a 3rd-order tensor model was made by using the user query to produce a concept window wherein terms are selected and run through the concept space using the TFDIF equation, in order to produce a concept vector for those concepts in the concept space (Hong et al., 2016). The research team identified that while this approach produced a list with high accuracy and correctness in terms of similarity function value, the 3rd-order tensor model can be much more time efficient and compact. The efficiency is achieved by introducing a model pool and by removing irrelevant documents among the concepts in the selected model.

1.1 Statement of the Problem

The research team aims to further decrease the size of the concepts by applying the similarity measure of the query to each of the concept in the model and filter out irrelevant concepts on these documents in the process. In this way, the search algorithm can theoretically run faster as less terms or concepts are being processed.

The study would look into these problems:

1. What is the actual amount of running time of the researchers' study on documents related to Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search technologies, as compared to the actual time of the existing study?
2. What is the accuracy of the retrieved documents as compared to the accuracy of the approach of the existing study by Hong and Kim?

1.2 Significance of the Study

This study aims to compare the search accuracy and improve the response time of the current semantic search engine. This study is a significant endeavor for students who would like to seek for documents regarding the domain on technology-related documents specific to Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search; it may also be beneficial for students, researchers, technologists and professors who can use it as an effective way of looking for certain techniques and functions relating to these technology concepts in order to conduct their own researches and studies. This study can be used as a reference and as a basis for analysis and comparison for their future works.

2. RELATED STUDIES

2.1 Semantic Search

Merriam-Webster dictionary define semantics as meaning or relationship of meanings of a sign or set of signs. Applied to search, "semantics" essentially relates to the study of words and their logic. Semantic search seeks to improve search accuracy by understanding a searcher's intent through contextual meaning (Sander, 2016). According to Amerlands, semantic search is like a searchlight because it picks up all the different data nodes of the Web and follows them around creating a picture of how they link up, who they belong to, who created them, what else they created, who they are, who they were, and what they do (Amerland, 2014).

Teodora Petkova defined semantic search as a technique that, "reaches out beyond keywords and seeks to understand the semantics of the search query. It improves search accuracy by looking at both data and their connections". She also added that, "instead of more links, which are only a single kind of relation, the algorithm presents you with a networked view of relations, facts, information, you might not know even existed" (Petkova, 2016).

2.2 Information Retrieval

As defined by the Cambridge Press, information retrieval or IR is finding material, usually documents, of an unstructured nature, which are usually text that satisfies an information need from within large collections that are usually stored on computers (Manning et. al., 2006). Information retrieval is basically concerned with the retrieval of relevant information or documents from data stores or databases (Tan et. al., n.d.). IR is concerned with facilitating the access of a user to huge amounts of information, which are predominantly textual in nature (Tan et. al, n.d.). The documents can be books, journals, reports, or other records of thought, or any parts of such records; namely articles, chapters, sections, tables, diagrams, or even particular words. The retrieval devices can range from a bare list of contents to a large digital computer and its accessories. The range of the operations can be from simple visual scanning to the most detailed programming methods (Vickery, 1959).

A retrieval system can be studied at three levels:

1. First, the way in which units of information, and relevant relations between them, are defined in the system. This is the semantic level of subject analysis.

2. Second, the general structural features of the system considered as a network of units of information linked to each other and to documentary items. This may be called structural analysis.

3. Lastly, the physical mechanisms (hardware) in which the structure is embodied (Vickery, 1959).

2.3 Lesk Algorithm

In 1986, Michael E. Lesk introduced an algorithm for dealing with word sense disambiguation called Lesk Algorithm, which has then been considered a classical algorithm. This algorithm is based on the assumption that words that were provided in a given block of text will tend to share a common general topic (WSD, 2015). A simplified version of the Lesk algorithm is produced to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Versions have been adapted to use WordNet (WSD, 2015).

For example, given an ambiguous word and the context in which the word occurs, the Lesk algorithm returns a Synset (synonyms set) with the highest number of overlapping words between the context sentence and different definitions from each Synset.

Unfortunately, this method of word disambiguating is very sensitive to the exact wording of definitions, so the absence of a certain word can radically change the results of the query. In addition, Lesk's algorithm determines overlaps only among the glosses of the senses being considered. This is a significant limitation in that dictionary glosses tend to be fairly short and do not provide sufficient vocabulary to relate fine-grained sense distinctions (WSD, 2015).

Currently, more works appeared and are underway which offer different modifications of this algorithm. These works use other resources for analysis such as thesauruses, synonym dictionaries or morphological and syntactic models; For instance, it may use such information as synonyms, different derivatives, or words from definitions of words from definitions (WSD, 2015).

2.4 TF-IDF (Term Frequency-Inverse Document Frequency)

In information retrieval, TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. This statistic is often used as a weighting factor in information retrieval, text mining, and user modeling. Its value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Up to today, TF-IDF is one of the most well-known term-weighting schemes.

Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. It can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

In the equation below, based on the Apache Lucene Documentation, TF correlates to the term's frequency, defined as the number of times the term appears in the currently scored document.

$$TF = \sqrt{termFreq}$$

IDF's value correlates to the inverse of the number of documents in which the term appears. It means that rarer items give higher contribution to the total score. IDF appears for both the query and the document, therefore it is squared in the equation presented below.

$$IDF = 1 + \log(\frac{number\ of\ documents}{docFreq + 1})$$

In the equation below, the values of TF and IDF are multiplied with each other to complete the TF-IDF value.

$$TF - IDF = TF * IDF$$

2.6 Cosine Similarity Function

The similarity function is one of the most crucial algorithms used in the implemented study. As defined by Pennsylvania State University, "the similarity functions, also called similarity or distance measures, are essential to solve many pattern recognition problems such as classification and clustering". Similarity measure is a numerical measure of how alike two data objects are and often uses 1 and 0 to indicate whether a complete similarity or no similarity at all. According to Ashby, "The concept of similarity has a fundamental importance to almost every scientific field. Topological methods are applied in fields such as semantics. Graph theory is widely used for assessing cladistic similarities in taxonomy" (Ashby et. al., 2007).

As stated by Hong and Kim, "In the traditional Vector Space Model, documents are represented as term vectors and their similarity is generally measured by calculating the cosine similarity." The study used the Cosine Similarity Measure to define the similarity function between two term-by-concept matrices of documents.

In the equation below, d is equal to the term-by-concept matrix while q represents the query matrix.

$$\mathrm{sim}(\mathbb{d}, \mathbb{q}) = \frac{<\mathbb{d},\mathbb{q}>_F}{||\mathbb{d}||_2 \cdot ||\mathbb{q}||_2}$$

<d, q> denotes the Frobenius product in the equation presented below, which is equal to the trace of the matrices d and q.

$$< \mathbb{d}, \mathbb{q} >_F = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{d}_{ij} \cdot \mathbb{q}_{ij}$$

In the third equation below, $||d||_2$ denoted the L2-norm of matrix *d,* is computed.

$$||\mathbb{d}||_2 = \sqrt{< \mathbb{d}, \mathbb{d} >_F}$$

3. METHODOLOGY

The researchers believe that the introduction of similarity function on the concept vector of the 3rd-order tensor model can further improve the search time taken by the computation of the search algorithm from the previous model. If it is deemed possible, a search engine can accurately search and gather documents through information retrieval with lesser time whilst not affecting the accuracy and precision of the current model. In order to test this claim, the researchers conducted an experiment. There are 5 steps that make up this research project:

1. Researching about the basics of semantic search and information retrieval, as well as existing studies regarding semantic search engines. Find a model search engine and find ways to improve it.

2. Developing new approaches which can improve the accuracy and speed of this existing semantic search engine, therefore enhancing its overall performance.

3. Programming/Coding the new implemented approaches and implementing it into the existing architecture in order to build a new implemented system.

4. Produce adequate test cases to examine the new implemented system.

5. Analysis of the results on whether there is improvement between the model system and the new system.

In the first step, the researchers conduct investigation and research works regarding the concepts involved such as semantic search and information retrieval. Knowledge on past researches were also gathered and different approaches were taken into consideration in this phase. In the second step, the researchers converge all of the gathered information and find the pros and cons of these models. If improvement on a model is detected, a new approach is considered for that system. This new approach is then made into a pseudocode, ready for programming phase. In the third step, the approach's pseudocode is converted into a program, as well as the user interface which end users will use. In the fourth step, multiple test cases that cover several scenarios of input into the system are conducted to test the performance of the new approach that is being introduced. In the fifth and final step, the results are analyzed and weighed using the evaluation tools (precision, recall, and F1-measure) in order to know whether the system fares better than the existing system.

The training data in this study is gathered using random sampling. The documents are in PDF form and are limited on technology related documents specific to Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search that are only written in English. The total number of documents needed from the IEEE database is 250 and these were retrieved from November 13, 2017 to November 20, 2017, while those gathered from Wikipedia are approximately 15,000 and these were retrieved from September 11, 2017 to November 9, 2017.

The researchers created two sets of queries, namely, Set A, and Set B. Each set is divided into five categories, namely; Data Compression, Image Processing, Information Extraction, Semantic Search, and Word Disambiguation. The categories contain five queries each. Each category is described as follows: 1) the first query is the definition of each particular category, 2) the second query contains a query that is highly related to the category, 3) the third query contains a query that is good or related to the category, 4) the fourth query comprise of a query that is somewhat related to the category, and 5) the fifth query contains a query that is not related to the category. There are two sets used to test the data, namely; Set A is used to compare the actual running time of the implemented system with the sorted concepts and pruning feature compared to the original system. Table 1 below shows the list of queries for Set A that were tested.

Table 1: List of Queries for Set A

Category	ID	Queries
Word disambiguation	Q1	determine the proper sense of an ambiguous word in a given context
	Q2	the task of determining the correct sense of a word in context
	Q3	the correct meaning of a word
	Q4	natural classification problem
	Q5	Word-sense induction
Data compression	Q1	the process of reducing the amount of data needed for the storage or transmission of a given piece of information
	Q2	compressing data into smaller files
	Q3	manipulating order of data
	Q4	extraction of data
	Q5	computer programming
Image processing	Q1	the analysis and manipulation of a digitized image, especially in order to improve its quality
	Q2	processing images or pictures
	Q3	changing the picture's data
	Q4	image manipulation
	Q5	getiing data from music
Information extraction	Q1	the process of turning the unstructured information embedded in texts into structured data
	Q2	automatically extract structured information
	Q3	extract fields of interest from free text
	Q4	text simplification technique
	Q5	data gathering
Semantic search	Q1	a data searching technique in which a search query aims to not only find keywords, but to determine the intent and the context of the search
	Q2	information searching with context and intent
	Q3	data searching context
	Q4	searching technique
	Q5	microsoft windows

Set B is used to compare the actual running time between the implemented system with result pool and without the result pool. Table 2 below shows the list of queries for Set B that were tested.

Table 2: List of Queries for Set B

Category	ID	Queries
Word disambiguation	Q1	determine the proper sense of an ambiguous word in a given context
	Q2	the task of determining the correct sense of a word in context
	Q3	the correct meaning of a word
	Q4	natural classification problem
	Q5	word meaning induction
Data compression	Q1	the process of reducing the amount of data needed for the storage or transmission of a given piece of information
	Q2	compressing data into smaller files
	Q3	manipulating order of data
	Q4	extraction of data
	Q5	computer programming
Image processing	Q1	the analysis and manipulation of a digitized image, especially in order to improve its quality
	Q2	processing images or pictures
	Q3	changing the picture's data
	Q4	image manipulation
	Q5	getiing data from music
Information extraction	Q1	the process of turning the unstructured information embedded in texts into structured data
	Q2	automatically extract structured information
	Q3	extract fields of interest from free text
	Q4	text simplification technique
	Q5	data gathering
Semantic search	Q1	a data searching technique in which a search query aims to not only find keywords, but to determine the intent and the context of the search
	Q2	information searching with context and intent
	Q3	data searching context
	Q4	searching technique
	Q5	Windows operating system

To determine if the accuracy of the implemented system is maintained, the researchers used same Set A queries to both the original and the implemented system. The accuracy of the implemented system is then compared to the accuracy of the original system. As for Set B, the queries are used to the implemented system and the accuracy is then compared to the accuracy when using the result set of Set A.

On identifying whether a document is relevant to the query or not, the researchers used the Apache Lucene search engine. Each query is used in the Apache Lucene search engine and the resulting result set are then used as the set of relevant document for that particular query.

For the evaluation on the accuracy of the results given from the system, the researchers use the precision, recall and F – measure. Precision measurement is the measure of the percentage of retrieved documents that are relevant. Recall is the percentage of relevant documents retrieved. F1 – measure is a measure that combines precision and recall using the harmonic mean of precision and recall.

In measuring the precision and recall, the number of True Positives, False Positives and False Negatives are needed to be known. The True Positives are the set of documents that are correctly retrieved from the set of documents which are described to be “correct”. The False Positives are the set of documents which are not supposed to be retrieved but retrieve. The False Negatives are the set of documents which are supposed to be retrieved but are not returned.

The precision measurement computed below show that T_p is the number of true positives and F_p is the number of false positives.

$$\boldsymbol{P = \frac{T_p}{T_p + F_p}}$$

The recall measurement computed below show that F_n is the number of false negatives.

$$\boldsymbol{R = \frac{T_p}{T_p + F_n}}$$

In the equation below, the F1-measure is computed using the precision and recall values, computed using the equations presented above.

$$\boldsymbol{F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}}$$

4. Analysis and Interpretation

4.1 The Implemented System

The system processing is divided into two (2) phases: the creation phase and search phase.

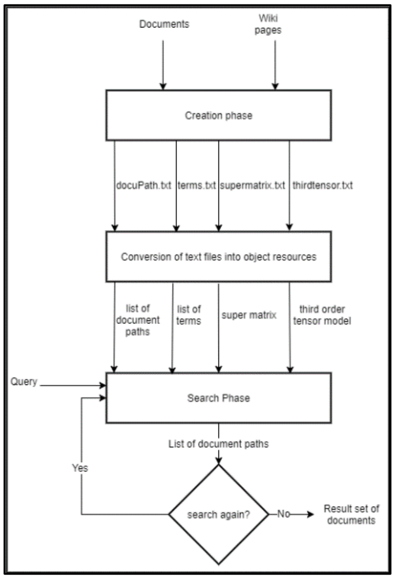

Figure 1: System Architecture of the Implemented System

The creation phase is the phase that creates resources for the search phase depending on the documents and Wikipedia pages as inputs of the phase. This phase creates the resources only once. However, if changes are applied to the documents or the Wikipedia pages, then it is necessary to run this phase again before using the search phase. The search phase is the phase where it uses the query from the user and the resources extracted from the “Conversion of Text Files into Objects” module to generate the result set and present it back to the user. The “Conversion of Text Files into Objects” module is a module that is run once to prepare the resources needed for the search phase. However, due to the web application technology in Java, the first request to the web application runs the “Conversion of Text Files into Objects” module. In other words, the first request in the web application has an overhead of extracting resources generated from the creation phase.

The search phase accepts queries from the user and preprocesses it. After that, it passes the preprocessed query to the Lucene Search module where it produces a list of concepts related to the query. This list of concepts of the user query concept space is used to refine the 3rd – order tensor model and Super Matrix by using the Refinement module. The outputs of the Refinement module are the Refined 3rd – order tensor model and the Refined Super Matrix. The Refined 3rd - order tensor model is used later in the Search module if the result history module does not find a query similar to the current query being searched.

The Refined Super Matrix and the preprocessed query are passed to the Query Matrix module and the output of this module is a Query Matrix for the current query. This Query Matrix and the threshold is used in the Result History module to search similar query in the result set pool and decide whether to select a record in the pool or search through the tensor model.

If a match was found in the pool, then the result set of document of the matched query is passed to the output view where it will display each documents. However, if a match

was not found, then the Refined 3rd - order tensor model, produced by the Refinement module, is used along with the Query Matrix, constructed from the Query Matrix module, in the Search module.
The result set of documents searched by the Search module is stored in the result set pool along with the current query and the Query Matrix. After storing the resources in the result set pool, the result set of documents is passed to the output view to present each documents to the user.

4.2 Test Results

In order to test whether the study affects the actual running time in any possible way of the implemented search engine by the researchers, the researchers prepared two sets of queries, which are Set A and Set B respectively, that were defined in Chapter III of the documentation.

The first set, Set A, is used to compare the actual running time of the implemented system with the sorted concepts and pruning feature compared to the original semantic search system by Kim Han-Joon and Hong Ki-Joo.

Set B is used to compare the actual running time between the implemented system by the researchers with result pool and the implemented system also but without the result pool.

Tables 3 and 4 below shows the results of the testing done on both Set A and B.

Table 3: Comparison of Actual Running Time for Set A (Original and Implemented)

Category	Number of Queries	Average of Original (sec)	Average of Implemented (sec)
Data Compression	5	11.571	6.2424
Image Processing	5	11.6222	6.5954
Information Extraction	5	12.075	8.273
Semantic Search	5	12.292	5.1166
Word Disambiguation	5	11.8522	6.5734
Average Actual Running Time		11.88248	6.56016

Table 4: Comparison of Actual Running Time for Set B (With RP and Without RP)

Category	Number of Queries	Average of Original (sec)	Average of Implemented (sec)
Data Compression	5	10.4094	3.5546
Image Processing	5	9.4356	3.5258
Information Extraction	5	11.769	3.6256
Semantic Search	5	9.7014	3.4904
Word Disambiguation	5	9.364	3.5402
Average Actual Running Time		10.13588	3.54732

4.3 Analysis of Results

The tests conducted by the researchers show that the speed in actual running time of the search by the new system for all the queries is significantly better and is not generated by chance. In addition, for the all components test case, the paired means t-test at 95% level has shown that new system is significantly better.

Based on the results presented in Tables 3 and 4 above, it can be deduced that the new system is better than the new system in terms of the speed in actual running time by 30%, its results are not generated by chance and is significant as shown in the paired means t-testing results in the documentation. The actual running time of implemented system also changes variably in every trial or search and the actual running time of current search engine always stays in 11 seconds. This is because the current search engine always uses the full dimensions of the model while the implemented system uses the size of the user query concept space to limit process of searching in using the whole dimensions.

5. CONCLUSIONS

The accuracy of the current system and the system with the refinement feature are the same but in terms of efficiency, the system with the refinement feature is faster than the current system.
However, the accuracy of the system with the result history feature is lower than the system with refinement feature and in terms of efficiency, the actual running time in seconds of the system with result history feature depends on the size of records stored in the pool.

6. RECOMMENDATIONS

If further improvements are to be made on this research study, the researchers would like to recommend the following notes and opinions:

1. Instead of utilizing the metric used by the proponents, use other third party metrics such as Semilar.
2. Provide adjustments on the algorithms, either by simplifying or enhancing, in order to acquire better speed and accuracy in the search process.
3. Provide adjustments on the algorithms being utilized in order to acquire better results on the evaluation tools used such as precision and recall, along with the F1-measure.
4. Utilize a more specific algorithm other than that used by the proponents.
5. Take into consideration other file types other than PDF files.
6. Consider including sentences with special symbols such as hyphens, commas, semicolons, etc.
7. Limit database by finding a metric that chooses database records to remove.

7. ACKNOWLEDGMENTS

We are grateful for the support of the UST IICS faculty and staff including Mr. Cecil Jose Delfinado, Ms. Janette Sideño, Mrs. Perla Cosme, and to our Thesis Adviser, Mrs. Charmaine Ponay, who provided valuable and intellectual insights and opinions for the completion of our study.

We are also thanking our peers and acquaintances, for their assistance and accompaniment in favorable and trying times. We also owe our thanks also to our ever supportive parents and family members, for providing for our needs and understanding our endeavors.

Lastly, we would like to extend our earnest gratitude to Almighty God, for bestowing upon us wisdom and intelligence to complete this work.

8. REFERENCES

Amerland, D. (2014). *Google™ Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact, and Amplify Your Online Presence, 1st Ed*. Indiana: Pearson Education. doi:0789751348

Ashby, F. G., & Ennis, D. M. (2007). Similarity measures. Retrieved April 04, 2017, from http://www.scholarpedia.org/article/Similarity_measures,doi:10.4249/scholarpedia.4116

Banarjee, S. (2002). *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*. Duluth, Minnesota. University of Minnesota. DOI: 55812

Chowdhury, G. (2010). Basic concepts of information retrieval system [Abstract]. *Introduction to modern information retrieval*. doi:9781856046947

Dubin, D., & Smith, L. (2000). Major Events in the History of Information Retrieval. Retrieved April 04, 2017, from http://courseweb.ischool.illinois.edu/~kmedina/Lis329/IRHistory.html

Edwards, S. (2015). The Rapid Evolution of Semantic Search. Retrieved March 2, 2017, from http://www.inc.com/samuel-edwards/the-rapid-evolution-of-semantic-search.html

Fatima, A., Luca, C., & Wilson, G. (2014). *New Framework for Semantic Search Engine* (Unpublished master's thesis). Anglia Ruskin University. doi:10.1109/UKSim.2014.114

Feist, K. (2015). Compare Databases and Search Engines. Retrieved March 3, 2017, from http://www.library.illinois.edu/ugl/howdoi/compare1.html

Forsythe, G. (n.d.). History of Information Retrieval. Retrieved April 04, 2017, from https://www.asindexing.org/about-indexing/history-of-information-retrieval/

Herbirch, R., & Graepel, T. (2010). *Handbook of Natural Language Processing* (2nd ed.). Florida: CRC Press.

Hong, K., & Kim, H. (2016). *A semantic search technique with Wikipedia-based text representation model*, 177-182. doi: 10.1109/BIGCOMP.2016.7425818

Libut, R., Llaneta, M., Rey Lara, R., & Valencia, J. (2014). *Text - Based Project Document File (PDF) Search Engine Using XRank Algorithm* (Undergraduate). University of Santo Tomas.

Ma, S., Zhao, W., Zhang, S., & Zhang, H. (2012). *Material Hub: A Semantic Search Engine with Rule Reasoning* (Unpublished doctoral dissertation). Peking University. doi: 10.1109/COMPSACW.2012.17

Manning, C. D., Raghayan, P., & Schütze, H. (2009). Introduction to information retrieval. doi: 978 0 521 86571 5

Mulles, C., Pasion, G., See, S., & Sison, D. (2011). *What You Search Is What You Get: A Query - based Automatic Document Searching Using Semantic Search* (Undergraduate). University of Santo Tomas.

Orav, H., Fellbaum, C., & Vossen P. (2014). *Proceedings of the 7th International Global WordNet Conference (GWC 2014)*. pp. 78-85. Tartu, Estonia

Pennsylvania State University. (2014). 1(b).2.1: Measures of Similarity and Dissimilarity. Retrieved April 04, 2017, from https://onlinecourses.science.psu.edu/stat857/node/3

Petkova, T. (2016, May). Semantic Search: The Paradigm Shift from Results to Relationships. Retrieved March 2, 2017, from http://ontotext.com/semantic-search-the-paradigm-shift-from-results-to-relationships/

Petrovic, D. (2010). Semantic Web: Part of Business World. Retrieved March 2, 2017, from http://www.semanticweb.rs/Article.aspx?iddoc=32&id=65&lang=2

Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B., & Stefanescu, D. (2013, August). SEMILAR: The Semantic Similarity Toolkit. In ACL (Conference System Demonstrations) (pp. 163-168).

Sanders, A. (2016). What Is Semantic Search and What Should You Do About It? Retrieved March 1, 2017, from https://moz.com/blog/what-is-semantic-search

Sanderson, M., & Croft, W. (n.d.). *The History of Information Retrieval Research* [Pamphlet]. CIIR Publications.

Tan, C., & Sumanaweera, H. (n.d.). Information Retrieval. Retrieved March 29, 2017, from http://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol4/hks/inf_ret.html

Vickery, B. (1959). The Structure of Information Retrieval Systems. In *Proceedings of the International Conference on Scientific Information: Two Volumes* (Vol. 2, pp. 1275-1277). Washington D.C.: The National Academies Press. doi:https://doi.org/10.17226/10866

Wall, A. (2006). History of Search Engines: From 1945 to Google Today. Retrieved February 28, 2017, from http://www.searchenginehistory.com/

What is WordNet. (2015, March 17). Retrieved from https://wordnet.princeton.edu.

W.S.D. (2015). Retrieved April 04, 2017, from http://www.nltk.org/howto/wsd.html

Ydav, U., Narula, G., & Duhan, N. (2016). *A novel approach for precise search results retrieval based on semantic web technologies* (Unpublished doctoral dissertation). YMCA University of Science and Technology. doi: no doi available.

Yedidia, A. (2016, December). *Against the F-score.* Essay. Retrieved from https://adamyedidia.files.wordpress.com/2014/11/f_score.pdf.