University of Santo Tomas
Institute of Information and Computing Sciences

AN IR-BASED SEMANTIC SEARCH APPROACH ON PORTABLE DOCUMENT FORMAT (PDF) FILES USING SIMILARITY MEASURE

A Thesis Project

Proponents:
Agsunod, John Mark Robert M.
Banting, Carl Jayson M.
Brar, Harjit S.
Cunanan, Patrick Bryan F.

Thesis Advisor:
Ms. Charmaine S. Ponay

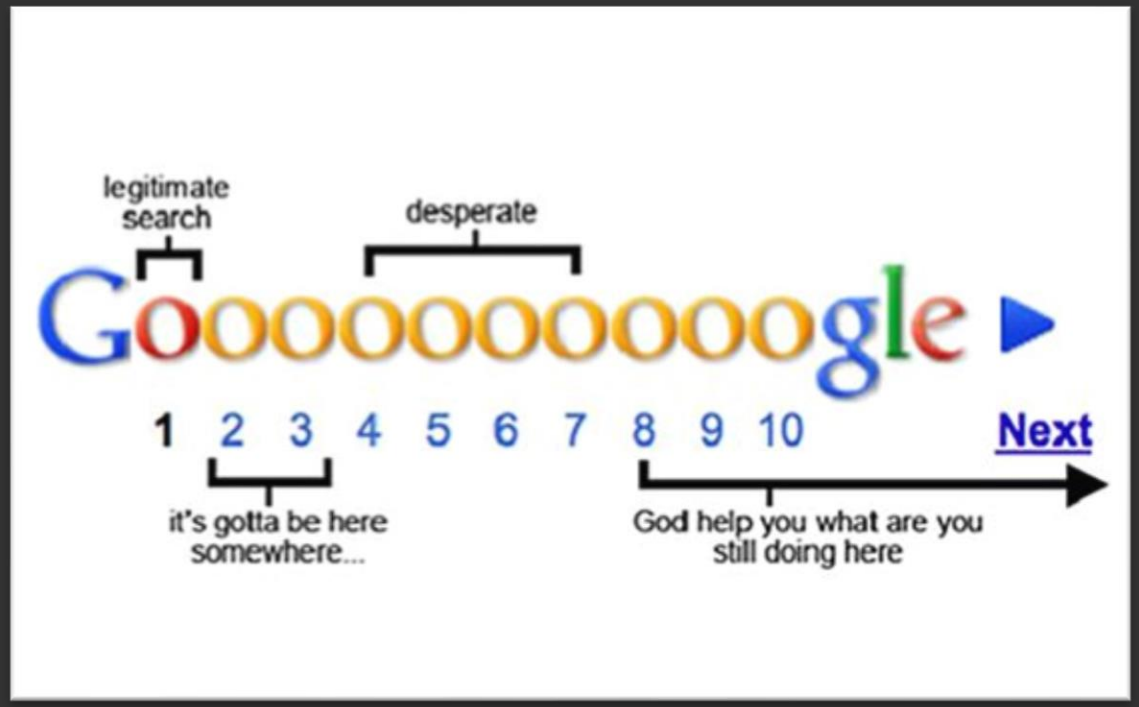
Thesis Coordinator:
Assoc. Prof. Perla Cosme

November 23, 2017

INTRODUCTION

- The idea of searching is changing.
- Then and Now
  - Tim-Berners Lee (1989)
- New Technology
  - AI
  - Neural Networks
  - Natural Language Searching

INTRODUCTION



In order to cope with this rapid data growth, techniques were created to find information that the user needed as swiftly and accurately as possible. One of the developed search techniques is *information retrieval.*

IR is a technique that mixes computer science and information science.

INTRODUCTION

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.

IR systems are designed to analyze, process and store sources of information and retrieve those that match the user's query.

INTRODUCTION

Current System

Ki-Joo Hong and Han-Joon Kim's *"A semantic search technique with Wikipedia-based text representation model"*

✓ *Aims to disambiguate terms by comparing terms with concepts*

PROBLEM: Efficiency. Uses all of the concepts.

Proposed System

BCAB's *"An IR-based Semantic Search approach on Portable Document Format (PDF) files using Similarity Measure"*

SOLUTION: Create result pool, reduced concepts by limiting

AIM: Faster system while maintaining the existing system's accuracy.

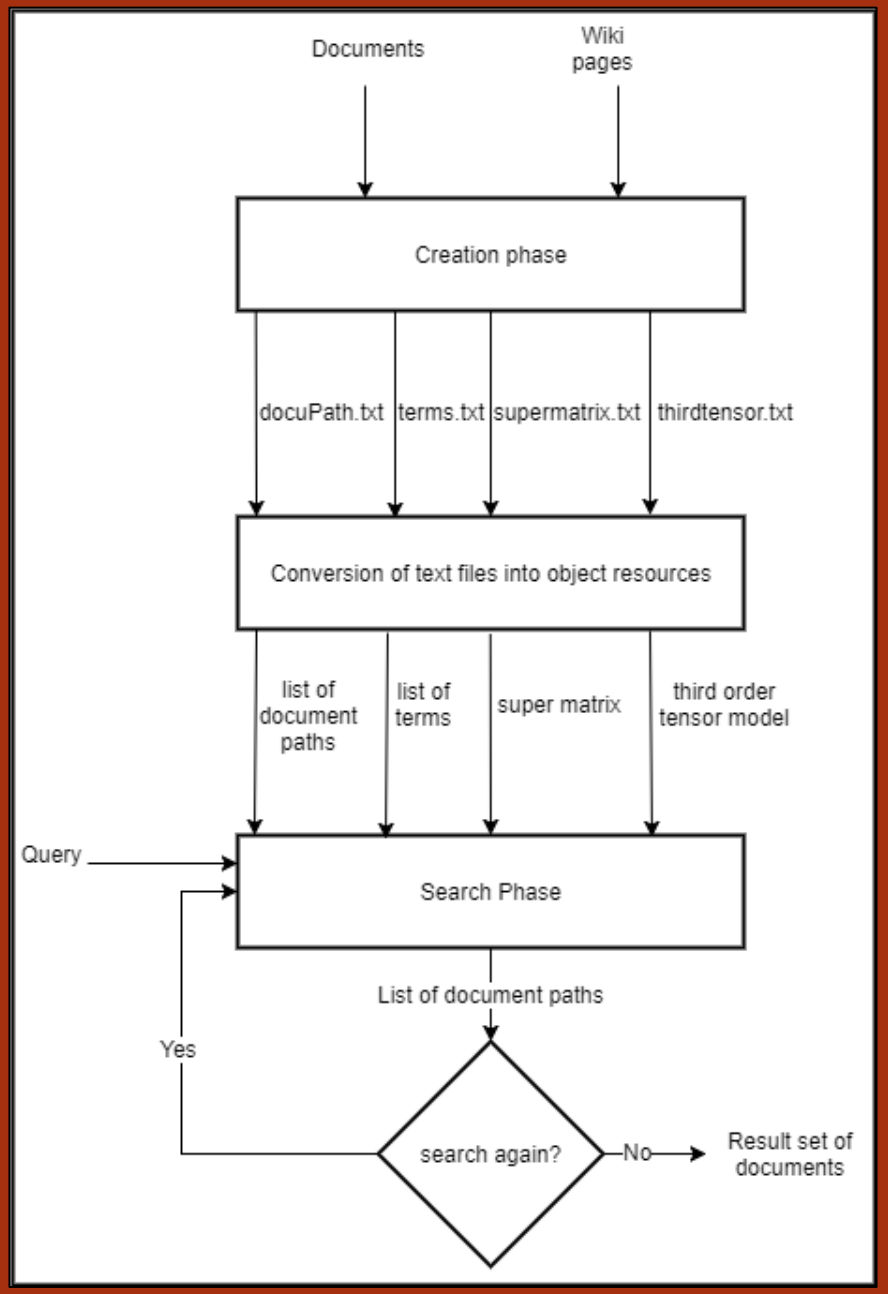
STATEMENT OF THE PROBLEM

The study would look into these problems:

- What is the actual time of the researchers' study on documents related to technology specific to Word Disambiguation, Information Extraction, Image Processing, Data Compression, and Semantic Search compared to the actual time of the existing study?
- What is the accuracy of the retrieved documents as compared to the accuracy of the approach of the existing study?

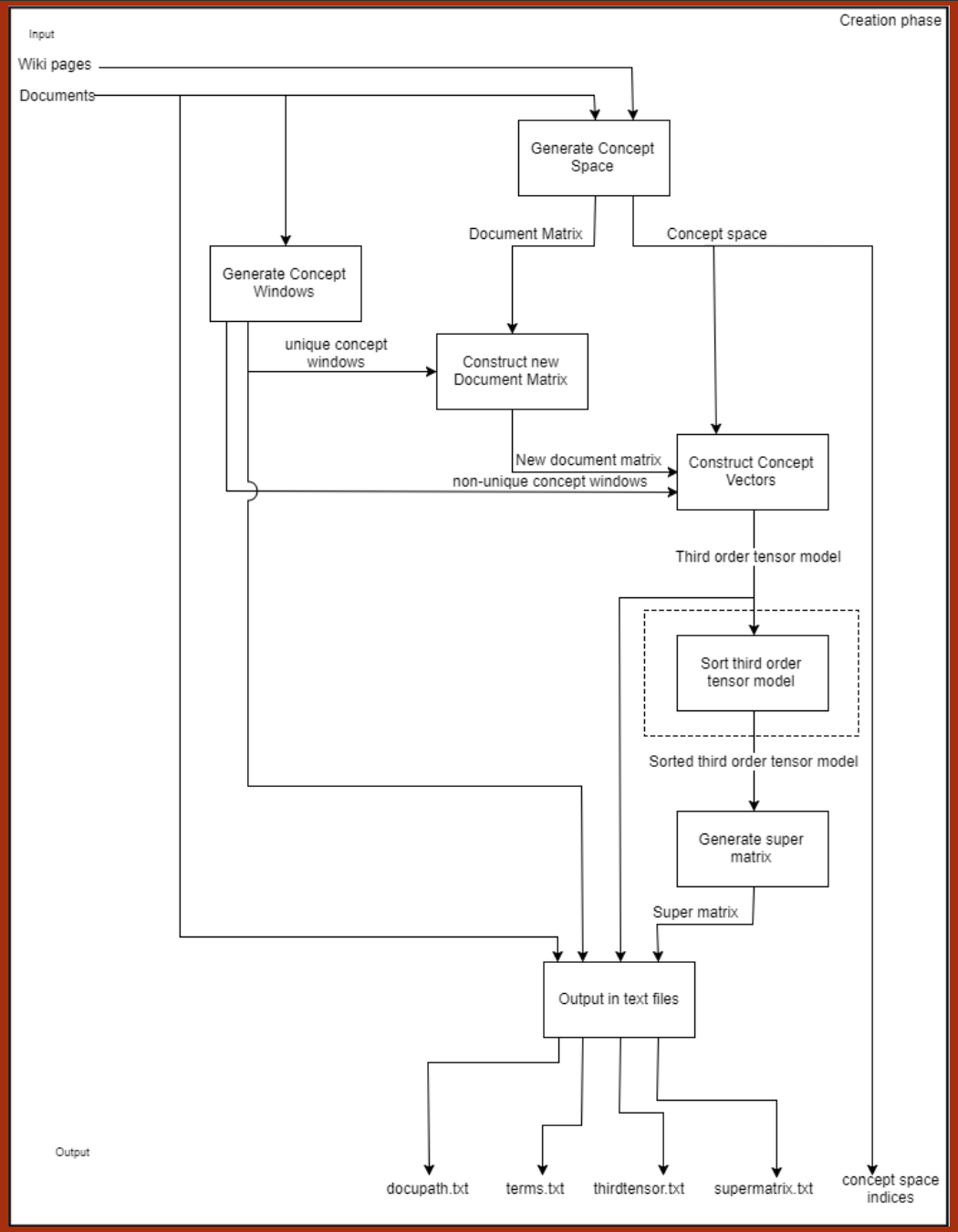
SYSTEM ARCHITECTURE

Creation and Search Phases



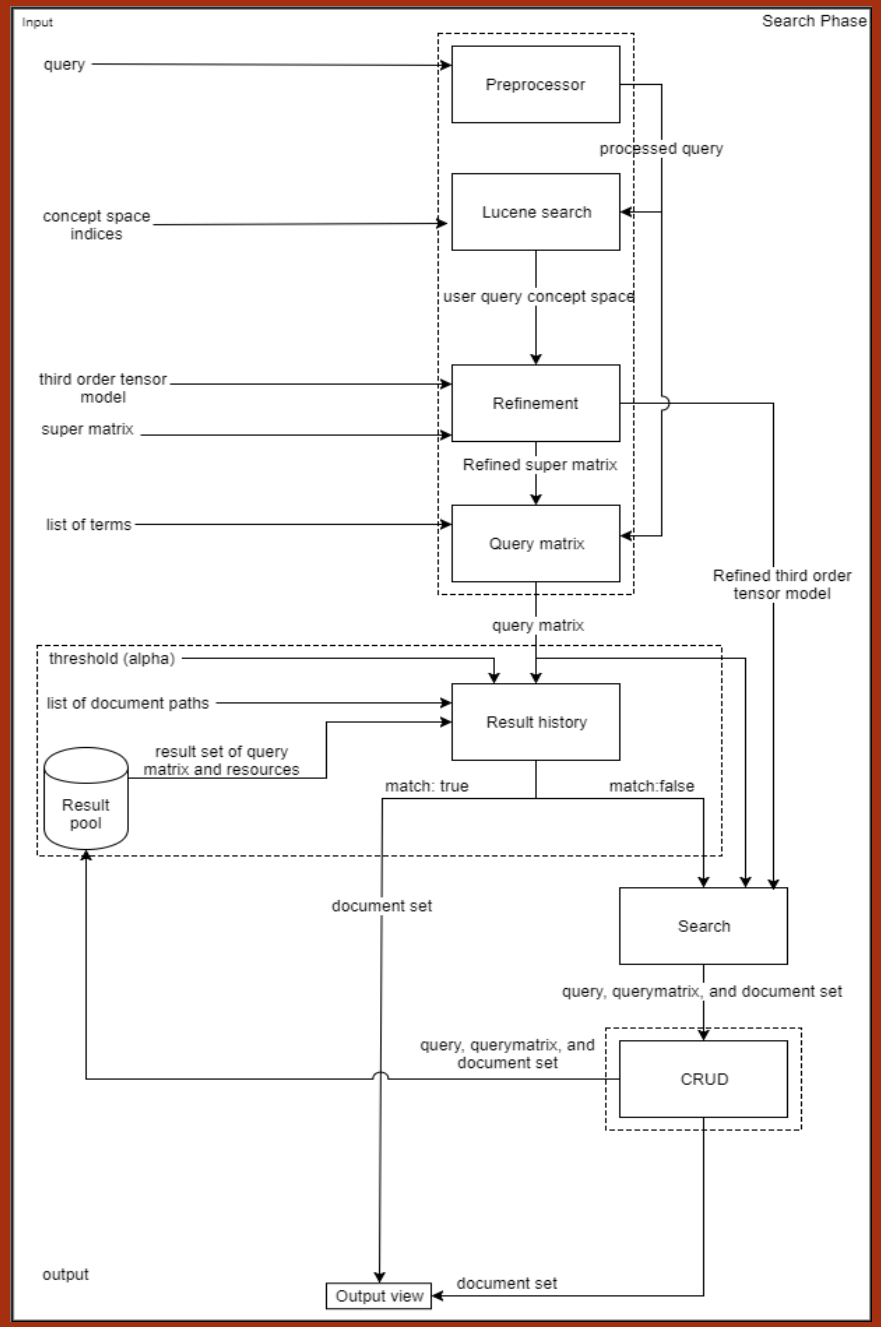
System Architecture of the System

CREATION PHASE



Level-2 System Architecture of the System (Creation Phase)

SEARCH PHASE



System Architecture of the System (Search Phase)

 SYSTEM DEMONSTRATION

PRESENTATION AND ANALYSIS OF DATA AND TEST RESULTS

Table IV-3. Comparison of Actual Running Time (Original and Implemented)

	Original (sec)	Implemented (sec)
Average Actual Running Time	11.882	6.56

Table IV-4. Comparison of Actual Running Time (With RP and Without RP)

	Without Result Pool(sec)	With Result Pool (sec)
Average Actual Running Time	10.136	3.547

Table IV-33. T-Statistic & Critical One-tailed Value (Set A)

	Set $A_{Original}$	Set $A_{implemented}$
Mean	11.883	6.560
Observations	25	25
Alpha	0.05	
t-Statistic	8.139	
Critical One-Tailed Value	1.711	

- Reject the null hypothesis
- Limiting the third tensor model also limits the process of searching through the model

- Reject the null hypothesis
- The implemented feature of limiting the 3rd-order tensor model and super matrix, is more efficient compared with the gathered running time of using Set B in a system, where it only has a feature of limiting or refining the 3rd-order tensor model and super matrix and the process of retrieving document set is done by searching through 3rd-order tensor model.

Table IV-34. T-Statistic & Critical One-tailed Value (Set B)

	Set $B_{with\ RP}$	Set $B_{without\ RP}$
Mean	10.136	3.547
Observations	25	25
Alpha	0.05	
t-Statistic	10.245	
Critical One-Tailed Value	1.711	

Table IV-35. Computation of Space Complexity for the Creation Phase

Module	Original	Implemented
Concept Space	$O((D \cdot T)+D+C)$	$O((D \cdot T)+D+C)$
Concept Window	$O(D \cdot (T \cdot 9))$	$O(D \cdot (T \cdot 9))$
Concept Vector	$O(D \cdot T \cdot C)$	$O(D \cdot T \cdot C)$
Tensor Sort	N/A	$O(C)$
Super Matrix	$O(D \cdot T \cdot C)$	$O(D \cdot T \cdot C)$
Total	$O((D \cdot T)+D+C)+O(D \cdot (T \cdot 9))$ $+O(2(D \cdot T \cdot C))$	$O((D \cdot T)+D+C)+O(D \cdot (T \cdot 9))$ $+O(C)+O(2(D \cdot T \cdot C))$

- The implemented system needs an addition $O(C)$ space.

Table IV-15. Average Precision, Recall, and F1-Measure of Set A

	Original			Implemented		
Category	Average Precision	Average Recall	Average F1-Measure	Average Precision	Average Recall	Average F1-Measure
Data Compression	0.98408	0.98408	0.98408	0.98408	0.98408	0.98408
Image Processing	0.95934	0.90557	0.92468	0.95934	0.90557	0.92468
Information Extraction	0.98719	0.88500	0.91794	0.98719	0.88500	0.91794
Semantic Search	0.99743	0.96502	0.97979	0.99743	0.96502	0.97979
Word Disambiguation	0.95025	0.90229	0.92270	0.95025	0.90229	0.92270
Overall Average	0.97566	0.92839	0.94584	0.97566	0.92839	0.94584

Table IV-37. MSE Scores for Precision scores (Set $A_{original}$ & Set $A_{implemented}$)

Category	Q_1	Q_2	Q_3	Q_4	Q_5
Data compression	0	0	0	0	0
Word disambiguation	0	0	0	0	0
Image processing	0	0	0	0	0
Information extraction	0	0	0	0	0
Semantic search	0	0	0	0	0

Table IV-38. MSE Scores for Recall scores (Set $A_{original}$ & Set $A_{implemented}$)

Category	Q_1	Q_2	Q_3	Q_4	Q_5
Data compression	0	0	0	0	0
Word disambiguation	0	0	0	0	0
Image processing	0	0	0	0	0
Information extraction	0	0	0	0	0
Semantic search	0	0	0	0	0

Table IV-39. MSE Scores for F1 scores (Set $A_{original}$ & Set $A_{implemented}$)

Category	Q_1	Q_2	Q_3	Q_4	Q_5
Data compression	0	0	0	0	0
Word disambiguation	0	0	0	0	0
Image processing	0	0	0	0	0
Information extraction	0	0	0	0	0
Semantic search	0	0	0	0	0

- Error using mean square error measurement,
  - $\varepsilon = 0\%$, for precision
  - $\varepsilon = 0\%$, for recall
  - $\varepsilon = 0\%$, for f1 score
- No improvement in accuracy
  - Indicates that the proposed algorithm of limiting the third tensor model does not affect the result
  - Model was arranged from highest to lowest

Table IV-28. Average Precision, Recall, and F1-Measure of Set B

	Set B			Set B'		
Category	**Average Precision**	**Average Recall**	**Average F1-Measure**	**Average Precision**	**Average Recall**	**Average F1-Measure**
Data Compression	0.99014	0.94112	0.96311	0.60233	0.60721	0.59889
Image Processing	0.92432	0.80303	0.82183	0.31042	0.40094	0.31711
Information Extraction	0.98276	0.88873	0.92223	0.73532	0.72756	0.70743
Semantic Search	0.99150	0.96564	0.97766	0.77824	0.69707	0.72741
Word Disambiguation	0.95766	0.95766	0.95766	0.77766	0.78315	0.77195
Total Average	0.96928	0.91124	0.92850	0.64079	0.64319	0.62456

Table IV-40. MSEs of Eval Scores from Queries (Data Compression

Scores	Mean square error (MSE)
Precision	0.1790
Recall	0.1656
F1	0.1733

Table IV-41. MSEs of Eval Scores from Queries (Image Processing)

Scores	Mean square error (MSE)
Precision	0.498828
Recall	0.369432
F1	0.416042

Table IV-42. MSEs of Eval Scores from Queries (Information Extraction)

Scores	Mean square error (MSE)
Precision	0.105199
Recall	0.134242
F1	0.079913

Table IV-43. MSEs of Eval Scores from Queries (Semantic Search)

Scores	Mean square error (MSE)
Precision	0.09134
Recall	0.164092
F1	0.137053

Table IV-44. MSEs of Eval Scores for Queries (Word Disambiguation)

Scores	Mean square error (MSE)
Precision	0.04968
Recall	0.041404
F1	0.043261

SUMMARY OF FINDINGS

The summary of findings discovered and analyzed from the previous section are as follows:

- The system with the refinement module offers same accuracy with the original system but it is more efficient in terms of actual running time.
- The system with the result history module offers low accuracy as compared to the original system where the searching through the 3rd-order tensor model is processed.
- The similarity metric and method used in the result history module depends on lexical matching. Thus, it cannot detect synonym words.
- The creation phase with the sorting module needs more space than the original system

CONCLUSION

- The accuracy of the proposed system **stayed the same**
- The actual time or speed is **decreased by 30%**
- The proposed system garnered a **mean square error of 0.0136 or 1.36%**

RECOMMENDATIONS

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

- Instead of utilizing the metric used by the proponents, use other third party metrics such as Semilar.
- Provide adjustments on the algorithms, either by simplifying or enhancing, in order to acquire better speed and accuracy in the search process.
- 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.
- Utilize a more specific algorithm other than that used by the proponents.
- Take into consideration other file types other than PDF files.
- Consider including sentences with special symbols such as hyphens, commas, semicolons, etc.
- Limit database by finding a metric that chooses database records to remove.