

Research Statement

Iman Rasekh, PhD
Computer Science, Information Retrieval

RESEARCH INTEREST

Information Retrieval, Web Scraping, Search Engine Optimization, Word Sense Disambiguation (WSD), Evolutionary Algorithms.

MY PhD RESEARCH SUMMARY

University of the Philippines at Los Banos (UPLB); Research Assistant (Information Retrieval)
2011-2017

- Developing Search Engine, MISE, based on a new four-layered core structure that is less complicated than the other existing core architectures for search engines such as Google or Yandex.
- Using a redefined version of the Imperialistic Competitive Algorithm (ICA) for large scale data clustering.
- Five common search engines (Google, Bing, Yandex, Yahoo and AOL) and 10 current page ranking algorithms (GPR, WPR, WLRank, TR, TIR, HITS, Clever, JBR, QDR and DRA) were compared to MISE in terms of precision score.
- A Simulation Experiment was done in seven steps and a t-test was performed.
 1. **Select the Test Environment:** Google Chrome was used as a common web browser
 2. **Select the Search Engines:** MISE, Google, Bing, Yandex, Yahoo, AOL and the developed environments for the current page ranking algorithms.
 3. **Choose the Searched Queries:** The 100 most-searched queries of the year 2015 in the Philippines were extracted from Google Trends and used as searched queries
 4. **Data Extraction:** The results of the search engines were extracted using SERP Scraping Tools
 5. **Delete the Priority URLs:** Priority websites (Wikipedia, Social Media Networks and Google-owned) were deleted from the results.
 6. **Calculate the Precision Score:** Precision score was used to judge the search results of each search engine
 7. **Empirical Comparison and the T-test:** A one-tailed t-test performed on average precision scores to compare MISE against common search engines and page ranking algorithms.
- A empirical comparison was performed using a one-tailed t-test and the results showed that MISE is capable of significantly better in fetching more relevant results than five commonly-used search engines: Google ($p=1.05E-05$), Bing ($p=0.0002$), Yandex ($p=0.0025$), Yahoo ($p=0.0006$), and

AOL ($p=4.1E-05$), and 10 well-known Page Ranking algorithms: General Page Ranking(GPR)algorithm ($p=6E-05$), Weighted Page Rank (WPR) algorithm ($p=0.0040$), Web page ranking Based on link attributes (WLRank) algorithm($p=0.0026$), Tag Ranking(TR) algorithm ($p=0.0038$), Time Rank (TIR) algorithm ($p=2E-05$), Hyperlink-Induced Topic Search(HITS) algorithm ($p=0.0086$), Clever algorithm($p=2E-05$), Journal based Ranking (JBR) algorithm ($p=0.0001$), Query Dependent Page Ranking (QDR) algorithm ($p=5E-05$), and Distance Rank (DRA) algorithm ($p=3E-06$).

- The average response time of MISE was empirically compared with the Google
- An Expert Evaluation was performed by experts to evaluate the results of the search engine
- **Achievements and contributions :**
 - **Design of a Content-Oriented Clustering Algorithm for Search Engines:** The modified version of ICA is capable of covering a large amount of data.
 - **Proposed a New Core Architecture for Search Engines:** This new architecture is less complicated than the common architectures and it was designed based on the architecture of Google.
 - **Implementation and Comparison with Current Page Ranking Algorithms:** can serve as an environment for further experiments and search engine comparisons.

FUTURE RESEARCH PLANS

1. Work on *Sense Detection* for semantic search engines
 - Work on Word Sense Disambiguation (WSD) as the most important part of sense detection, which is most challenging open problems in the field of information retrieval
 - Work on other types of sense detection like interrogative text detection using machine-learning algorithms, such as decision tree ,Perceptron or SVM (support vector machines)
2. Work on typographical errors in current search engines
3. Design a new WBIR (Web Based Information Retrieval system) that can cover different types of Mathematics formulas, media, search applications and tasks.

IMAN RASEKH