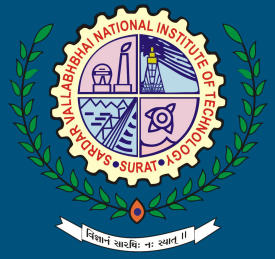
Research Paper Recommendation System (First Stage Project Presentation)



Prepared by

(Group 15)
Pankhi Khandelwal (U17CO099)
Himanshu Choudhary (U17CO101)
Shubhi Agarwal (U17CO109)
Aman Mishra (U17CO112)

Guided by: **Dr. Rupa G. Mehta**

Contents

Insight to the Problem

- Availability of infinite amount of data over web whose growth rate is exponential.
- Semantic comparison becomes almost infeasible.
- Searching out for relevant and related documents from such a huge corpus is a very tedious task.
- The problem also gets intensified when the users possess very skimmed knowledge of operating these digital repositories.
- Problem is not just restricted to research community but is slowly becoming part of any domain involving look-up from bulk data.
- Research paper recommendation aims to recommend new articles that match researchers' interests.

Problem Statement

- To build a graphical-network-model based on keywords, which correlates the input publication to its n-neighbors on the basis of their semantic relationships and associations.
- To use Natural Language Processing based techniques for filtering to reduce the search space for documents.
- To be followed by a personalized look-up in the search space and ultimately a semantic comparison in the resultant search space.

Literature Survey

- ➔ Research Paper Recommendation Approaches
- ➔ Research Paper Recommendation Systems and their Flow of Operation
- ➔ Keyword Extraction Techniques
- ➔ Similarity Measures

Research Paper Recommendation Approaches

- Citation Based
- Content Based
- Collaborative Filtering-Based
- Topic Based
- Keywords based
- Metadata Based

Research Paper Recommendation Systems and their Flow of Operation

- JournalFinder
- Google Scholar
- Springer

Keyword Extraction Techniques

- Simple Statistical Approaches
  - Word Frequency
  - Term Frequency Inverse Document Frequency (TF-IDF)
  - Rapid Automatic Keyword Extraction (RAKE)
- Linguistic Approaches
- Machine Learning Approaches

Similarity Measures

- Cosine Similarity
- Bibliographic Coupling
- Co-Citation
- Keyword Co-occurrence

Proposed Solution and Workflow

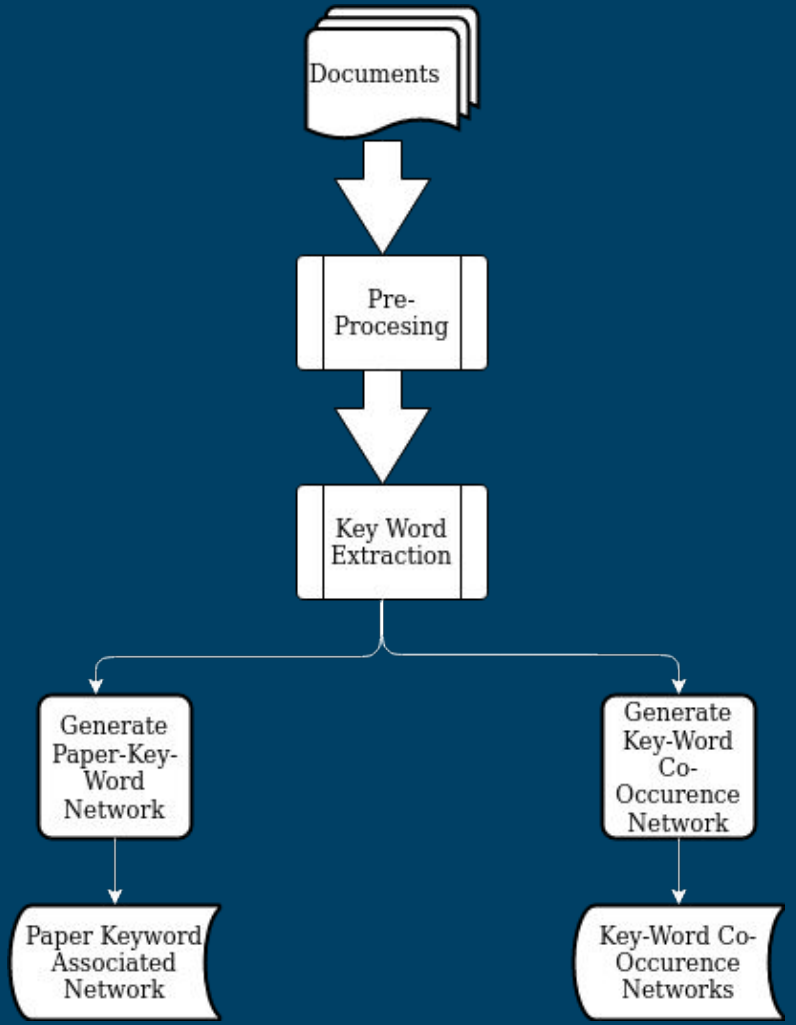
Makes use of a Graph based approach involving use of Keyword Co-Occurrence Network and Paper Keyword Associated Network.

Divided into two sections :

1. Model Preparation -> Covers the entire network creation
2. Model Application -> How user will interact with our system

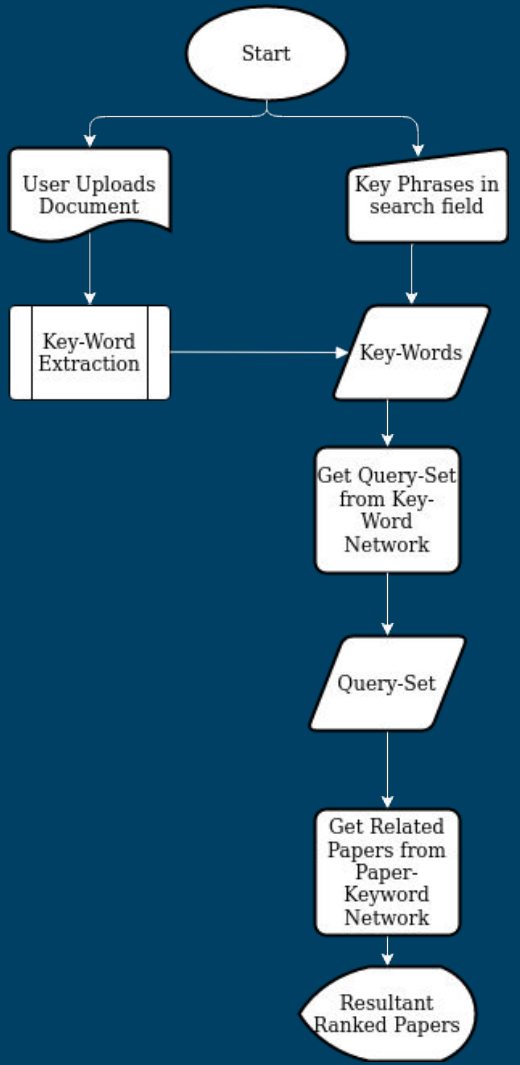
Model Preparation

(explains preprocessing and network creation stages)



Model Application

(explains how the user interacts with the system to get the recommendations)



How does the user interact with the system?

- **By Uploading a Research Paper:**
  Here the user can directly upload a research paper and as an output get all the research papers that are related to it. Internally, the key phrases will be extracted from the paper and the keyword-based network will be queried based on these keywords to get the desired results.

- **By Providing Key Phrase:**
  Here the user can provide key phrases. Based on these key phrases the Key Word network will be queried and as an output user will get the related papers.

How does the system work?

- From the user's query a set of all possible keywords will be generated (either provided by the user directly or extracted from the provided content) and fed to the system during initial stages.
- Further, the model prepares a query set by getting relatable keywords obtained for the input set of keywords or key phrases by using the context-information and semantic relationships extracted from keyword co-occurrence network.
- This query set is further utilized to get a set of connected and interlinked documents by analysing the paper-keyword network.
- From this collection of related papers, top n-ranked papers will be recommended to the user as an output.

Implementation and Results

- Pre-processing
  - Stemming
  - Lemmatization
  - Tokenization
- Keyword Extraction: RAKE Algorithm
- Network 1 (Paper IDs as nodes): Paper Keyword Network
- Network 2 (Keywords as nodes): Keyword Co-occurrence Network

Preprocessing Techniques

- **Stemming:** It is the process of extracting the base from the given words by removing affixes from them, this extracted part is known as stem. It reduces the given words to common stem.
  For example: good will be reduced to “good” and better will be reduced to “bett” in stemmer.

- **Lemmatization:** It is the process of extracting the valid base (root word not root stem) from the given words by removing affixes from them, this extracted part is known as lemma. It reduces the given words to common lemma.
  For example: good and better both will be reduced to “good” in Lemmatization.

  Github Link: https://github.com/Am-Coder/Document-Analysis/blob/master/preprocessing/preprocessing.py

- **Tokenization:** It is a process of splitting up a larger body of text into smaller lines, words, or even creating words for a non-English language.

  Github Link: https://github.com/Am-Coder/Document-Analysis/blob/master/preprocessing/tokenisation.py

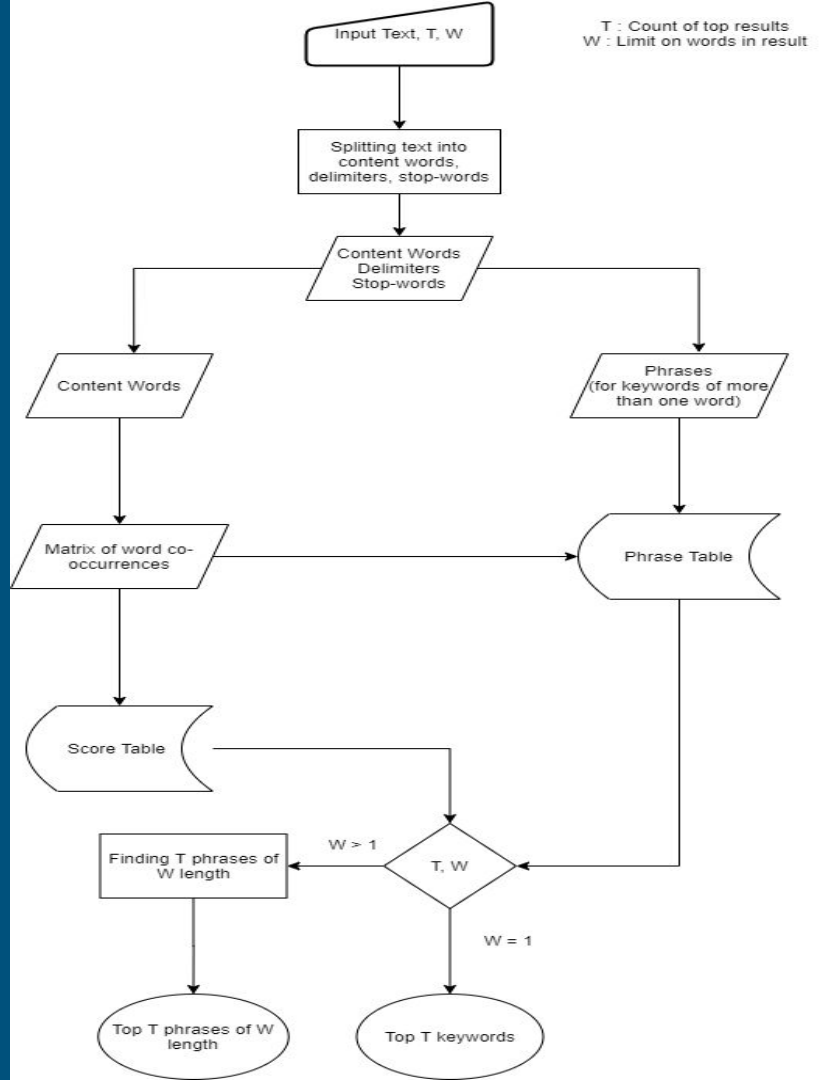
Keyword Extraction - RAKE

1. Splitting the text into content, delimiters and stop words.
2. Then the algorithm splits the text at phrase delimiters and stopwords to create candidate expression.
3. Now after extraction, the algorithm creates a matrix of word co-occurrences.
4. After matrix is built, words are given a score and score table is generated.

Score = degree of word / frequency of word

5. Returns the result as top T keywords of W words from score table.

Github Link:
https://github.com/Am-Coder/Document-Analysis/blob/master/keywordExtraction.py



Network 1 (Paper IDs as nodes): Paper Keyword Network

- A list of publication IDs for all the publications present in the dataset is generated in this system.
- Each element of the list acts as a node for an undirected graph.
- The weight of an edge is calculated by calculating the total number of common keywords between the two research papers depicted by the two nodes corresponding to that edge.

For example, suppose there are two scholarly articles with paper ids 1 and 2, list of keywords for both are as shown below:

Keywords for paper 1: [K1, K2, K3, K4, K5, K6]
Keywords for paper 2: [K1, K2, K4, K6, K7, K8]

So, the number of common keywords in both the articles is 4 (viz., K1, K2, K4 and K6), hence the weight of the connecting edge would be 4.

Github Link: https://github.com/Am-Coder/Document-Analysis/blob/master/Paper-KeyWord-Network/graphGeneratorFromKeywords.py

Network 2 (Keywords as nodes): Keyword Co-occurrence Network

- A list of common keywords extracted from all the publications present in the dataset is generated in this system.
- Each element of the list acts as a node for an undirected graph.
- The weight of an edge is calculated by calculating the co-occurrence frequency between the two nodes corresponding to that edge.

For example, suppose there are two keywords viz., K1 and K2, list of papers in which they are occurring ar as shown below:

Papers List for K1: [P1, P2, P3, P4, P5, P6]
Papers List for K2: [P1, P2, P6, P7, P8]

So, the number of articles having both K1 and K2 is 3 (viz., P1, P2 and P6), hence the weight of the connecting edge would be 3.

Github Link: https://github.com/Am-Coder/Document-Analysis/blob/master/Paper-KeyWord-Network/graphWithKeywordsAsNode.py

Applications

- A researcher in the current scenario has to manually search for the similar publications corresponding to his research to use them for preparing and studying the content related to his research or to perform the corresponding literature survey or to quote the citations.
- Therefore, a system that can automatically query an existing database to structure the co-related documents in a network and then look-up for the required one in the search space by reducing the search cost as compared to that of traditional means is of great significance and importance for practitioners.
- Along with researchers, any educational and research institution can also customize this module in order to provide an efficient e-library system for the members of their organization.

Summary

As it has been witnessed that the system for easy recommendation of scholarly articles is of great significance today due to the various quoted reasons, this project work would henceforth provide a probable solution to this demand. The main focus here is to develop a keyword-based recommendation system which also takes semantic relationships in consideration during the model development and utilization phases. Another worth noting fact is the capability of the model to reduce the search-space.

Future Prospects

- Further advancements in the project encompasses the extension of its domain by combining multiple approaches for article recommendation namely CF, content-based along with the semantic keyword approach as suggested.
- This whole flow aims to develop a personalised recommendation system which also takes into account users' data collected from their access log with the system.