MAJOR PROJECT PRESENTATION
ON
CONCEPT GRAPH EXTRACTION FROM NATURAL LANGUAGE TEXT

Submitted By:
Aman Kaushik, 11104656
Ujjval Dua, 11103611

Concept Map:

A Concept Map is a diagram that depicts suggested relationships between concepts. It is a graphical tool that is used to organize and structure knowledge. The technique for visualizing these relationships among different concepts is called concept mapping.

In a concept map, each word or phrase connects to another, and links back to the original idea, word, or phrase.

Concept maps are widely used in education and business. Uses include:

- Note taking and summarizing gleaning key concepts, their relationships and hierarchy
  from documents and source materials
- New knowledge creation: e.g., transforming tacit knowledge into an organizational resource, mapping team knowledge.
- Collaborative knowledge modeling and the transfer of expert knowledge.

Some relevant current/open problems:

Visualization and Understanding: Since a lot of information is available today relating to any topic, it becomes a tedious task to weed out the information relevant to ones query. To find this relevant information the user would have to read through pages and pages of texts.

Semantic search: Full text search engines find documents based on term frequencies, page ranks etc, instead of the concept inside and context of the user query.

Knowledge Discovery: Simple searches make it difficult for exploratory learning, research and ideation.

Problem Statement:

Extract concept maps out of unstructured natural language text efficiently and accurately. Build a system that can harness and apply them on any website, text document for easier understanding of the text, knowledge discovery, improving readability of a website, visually summarizing a text document for better and faster understanding a concept.

Overview of proposed solution approach and Novelty/benefits:

For this project the solution is divided into two separate tasks.

For Websites : Firstly, building a browser extensions that extracts text from the selected area on a website (A selected area could be represented by a block formed by dragging the mouse or by simply hovering the mouse over some part) passes it on to a server and displays the server response as a concept map of the selected text.

For Text Documents: Building a website in which a user can search for concept map relating to a particular subject, topic or chapter. The user would be able to fully interact with the map, i.e, partial viewing, highlighting nodes etc.

Secondly, a **concept extractor** which takes the text supplied by the browser extension as input. In first stage it resolves co references(both anaphora, cataphora, Split antecedents, Co-referring noun phrases) within the text. It then passes this resolved text to a binary relationship extractor, which then extracts <Concept, relationship, Concept> tuples from each sentence and stores it in a graph database. In the last stage these tuples are extracted from the graph database and sent to the browser extension/ website as response where an actual map is drawn from these tuples.

Benefit:

This system will help tremendously in improving the readability of a website as well as a document, and help users decide quickly whether the document contains the information he is looking for.

This would prove to be tremendously helpful for student as they would be able to visualize their theoretical knowledge and relate different concepts together.

Comparison with other existing approaches/ solution:

Other Solutions	Comparison
Machine Learning : Supervised Learning	This approach works best when the scope of application spans only across a particular domain. This approach doesn’t works as well as our approach for Wikipedia articles but performs better when applied to a particular domain such as a particular academic area.
Machine Learning : Unsupervised Learning	The accuracy of this approach is not limited by the domain and its also not partial to the domain. The performance of this approach with our approach is comparable but this on the other hand is more difficult and complicated to implement.
Fuzzy Logic	A direct comparison with this approach has not been done. But, the research with this approach is promising. This approach is also difficult to implement.

Details of Empirical Study:

New tools explored:

- **StanfordCoreNLP Parse:** Stanford CoreNLP provides a set of natural language analysis tools which can take raw text input and give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, and mark up the structure of sentences in terms of phrases and word dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, etc.

- **OLLIE Parser:** Ollie is software that automatically identifies and extracts binary relationships from English sentences. Ollie is designed for information extraction, where the target relations cannot be specified in advance.

- **Neo4j Graph Database:** Neo4j is an open-source graph database, implemented in Java. It is described as "embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables".

- **D3.js JavaScript Library:** D3.js is a JavaScript library for manipulating documents based on data. D3 helps bring data to life using HTML, SVG and CSS.

- **SpicyNodes:** SpicyNodes is a system for displaying hierarchical data, in which a focus node displays detailed information, and the surrounding nodes represent related information (Focus + Context), with a layout based on radial maps.

- **CmapTools:** The IHMC CmapTools program empowers users to construct, navigate, share and criticize knowledge models represented as concept maps.

- **Browser Extensions:** A browser extension is a computer program that extends the functionality of a web browser in some way.

Findings & Conclusion:

Findings

The results of this project support the following findings:

1. Summarizing text through concept graph visualization helps in faster understanding of the text and also helps in relating several different topics together.

2. Positive review were received for our results when shown to a few children. They claimed that it was easier to see the whole chapter as one through the concept maps.

3. Coreference resolution is the most crucial part of successful relationship extraction.

Conclusion

After testing our approach on Wikipedia articles, childrens' stories and academic textbooks it was seen that the approach works best when the content provide contains very little narrative. For eg: when extracting relationships out of stories, even though the relationships between primary characters were fairly accurate the flow of the story as whole was lost. Whereas on applying this approach on text books, which are mostly based on facts, this approach yielded very promising results.

We can say that extracting binary relationships from natural language text yields best results when the text doesn't contain any narrative.

Future Work

In future, we will extend this project by taking a supervised machine learning approach to get more accurate results. Another extension for this project would be to train our own model for NLP parsing, NER tagging and relationships extraction. Another approach to improve the efficiency can be to use Fuzzy Systems based approach.

References:

- Alberto J. Cañas, Roger Carff, Greg Hill, Marco Carvalho, Marco Arguedas,Thomas C. Eskridge, James Lott, Rodrigo Carvajal.”Concept Maps: Integrating Knowledge and Information Visualization.” Springer Lecture Notes in Computer Science, 2005.
- Dietrich Albert, Christina Steiner. “Representing Domain Knowledge by Concept Maps: How to Validate Them?”. 2nd Joint Workshop of Cognition and Learning Through Media-Communication for Advanced e-Learning (JWCL), 2005.
- Eugene Charniak and Micha Elsner, “EM Works for Pronoun Anaphora Resolution”
- Jorge J. Villalon, Rafael A. Calvo, “Concept Map Mining: A definition and a framework for its evaluation”
- Joseph D. Novak & Alberto J. Cañas, “The Theory Underlying Concept Maps and How to Construct and Use Them.”. Technical Report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition, 2008
- Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning, “A Multi-Pass Sieve for Coreference Resolution”
- Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni, “Open Language Learning for Information Extraction”
- Zubrinic, K., Kalpic, D., and Milicevic, M., “The automatic creation of concept maps from documents written using morphologically rich languages”