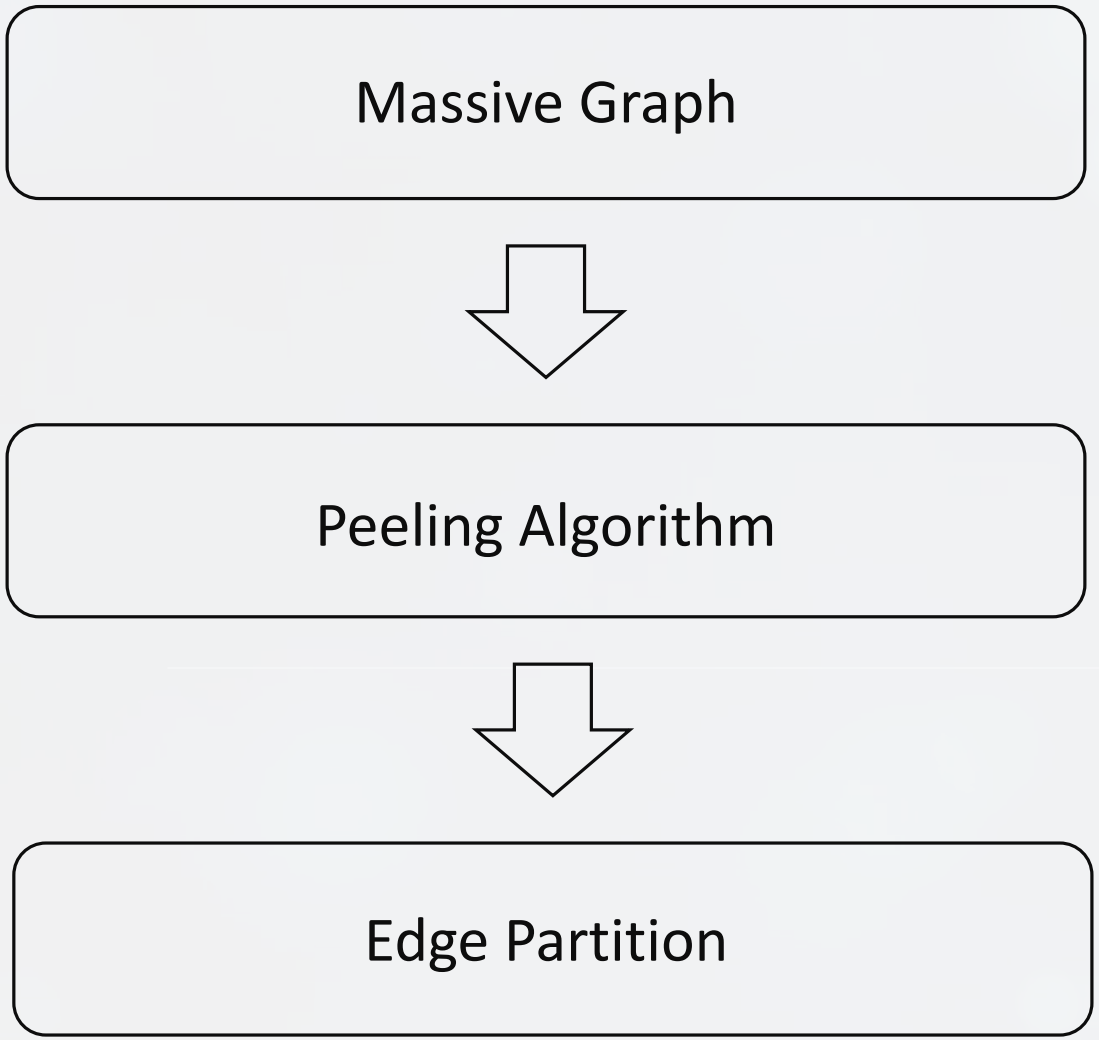
Graph Stories

Kevin Carman

Advised by Dr. James Abello

Recap



Goals

- Analyze the metadata of fixed points
- Algorithmically develop a ‘story’
- Integrate the algorithm into the Graph Waves project

Acquiring the data

- RCSB Protein Data Bank
  - Metadata - webpage
  - Scraped titles and abstracts of associated PubMed papers
  - Difficulty making connections

- US Patent Citation Network
  - Metadata – patent ID
  - Utilized Google Patents and PatentsView API
    - Patent title and abstract
  - Much easier to understand
  - Golf bag covers

Initial Approach

- AutoPhrase
  - Automated Phrase Mining
  - Training
    - 1GB file of relevant abstracts
  - Input – corpus
  - Output – corpus with tagged phrases
    - <phrase>example</phrase>
  - Thresholds

Initial Approach

- Phrase Selection
  - Parsed phrases from tagged corpus
  - Term Frequency vs. Inverse Document Frequency (TF-IDF)
  - Select Top-k phrases
    - $5 \leq k \leq 10$

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = \ln\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in them}}\right)$$

$$TF\text{-}IDF(t) = TF(t) * IDF(t)$$

Initial Approach

- Elasticsearch
  - Store, Search, and Analyze data
  - Indexing
    - 100MB file of relevant abstracts
    - Scraped data
  - Querying
    - Pairs of phrases
    - Pool of Top-j documents

Documents in pool = $\binom{k}{2} * j$

Initial Approach

- Sentence Selection
  - Find key sentences from pool
  - Best Matching 25 (BM25)
    - Ranking function for search engines
  - Greedy set cover algorithm

Initial Results

Example Abstract

A golf bag protective cover composed of plastic sheet material having slit openings therein which are covered by a plastic skirt which circumvents the entire cover, thus preventing dislodgment by wind while protecting the interior of the golf bag against ingress of rain.

Initial Phrases

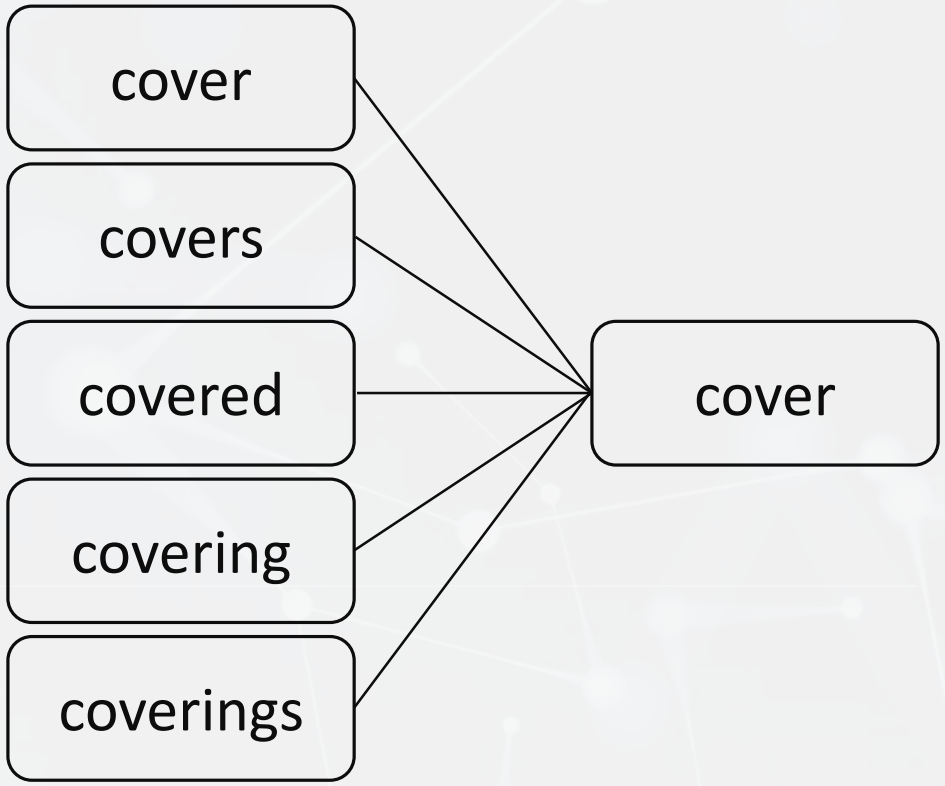
‘lower’, ‘fastener’, ‘invention’, ‘plastic’, ‘cover’, ‘covered’, ‘golf cart’, ‘upper’, ‘car’, ‘rain’

Initial Sentences

"The cover of this invention is preferably constructed of crushable or non-selfsupporting plastic so that it may be collapsed to a compact package for storage", "A front flap with front lateral side edges merges with the back lateral side edges, the front flap extending downwardly from the apex to a lower front edge, wherein the lower front edge of the front flap and a top edge of the front base wall define a flexible forwardly open mouth in an open configuration, and wherein the lower front edge of the front flap overlaps the top edge of the front base wall covering the mouth in a closed configuration.", "A quick-disconnect fastener is provided for attaching the access flap to the golf bag rain cover below the access slot."

Improvements

- Tweaked thresholds
- Bigger training sets
- Refactored and optimized
- Developed a workflow
- Phrase Selection
  - Minor data improvements
  - TF-IDF adjustments
    - Stemming
- Sentence Selection
  - Various algorithms tested
    - Venn diagram
  - Sentence ranking hierarchy



Improved Results

Improved Phrases

‘golf’, ‘lightweight’, ‘easy access’, ‘weather’, [‘cover’, ‘covered’], ‘play’, ‘rain’, ‘plastic’, ‘water’, ‘open configuration’

Improved Sentences

"A band of elastic surrounds and gathers the base and the top portions thereof in pleats, and the sleeve is sufficiently elongated so that when the base is fitted over the mouth of the bag, the upper portion of the cover will fold over the club heads.", "A flexible, waterproof, lightweight compact cover for the top of a golf bag and for golf clubs having a first end and a second end.", "The cover of this invention keeps clubs dry while in the bag, and, in addition, provides easy access and an unobstructed view thereof to facilitate selecting a club for play during such weather.", "A front flap with front lateral side edges merges with the back lateral side edges, the front flap extending downwardly from the apex to a lower front edge, wherein the lower front edge of the front flap and a top edge of the front base wall define a flexible forwardly open mouth in an open configuration, and wherein the lower front edge of the front flap overlaps the top edge of the front base wall covering the mouth in a closed configuration.", "A flexible golf bag cover for protecting the interior and contents of a golf bag of the type having an open top surrounded by a peripheral rim and having a carrying strap extending laterally from the rim.", "The cover (10) is composed of a clear, flexible, water impervious plastic, such as polyethylene.", "To gain access to the golf club, the second end of the cover is lifted and the golfer may reach in to remove the golf club while still preventing rain from entering the top of the golf bag."

Future Work

- Continuous improvement to Phrase and Sentence Selection
- Phrase Selection
  - word2vec
    - Train model
    - Mostly implemented
- Sentence Selection
  - Cluster similar phrases to cover a more diverse set of sentences
- Quantitatively evaluate progress
- Integrate the algorithm into the Graph Waves project

Acknowledgements

Special thanks to NSF Grant CCF-182215!

DIMACS
Dr. James Abello
Jingbo Shang
Haodong Zheng

References

- Atlas: Local Graph Exploration in a Global Context
- Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han, " Mining Quality Phrases from Massive Text Corpora" , accepted by IEEE Transactions on Knowledge and Data Engineering, Feb. 2018.
- Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren and Jiawei Han, "Mining Quality Phrases from Massive Text Corpora", Proc. of 2015 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'15), Melbourne, Australia, May 2015.
- Kai Hong, Ani Nenkova, "Improving the Estimation of Word Importance for News Multi-Document Summarization", Proc. of the 14th Conf. of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, April 2014.
- Martin Porter, "An algorithm for suffix stripping", *Program* 14 no. 3, pp 130-137, July 1980.