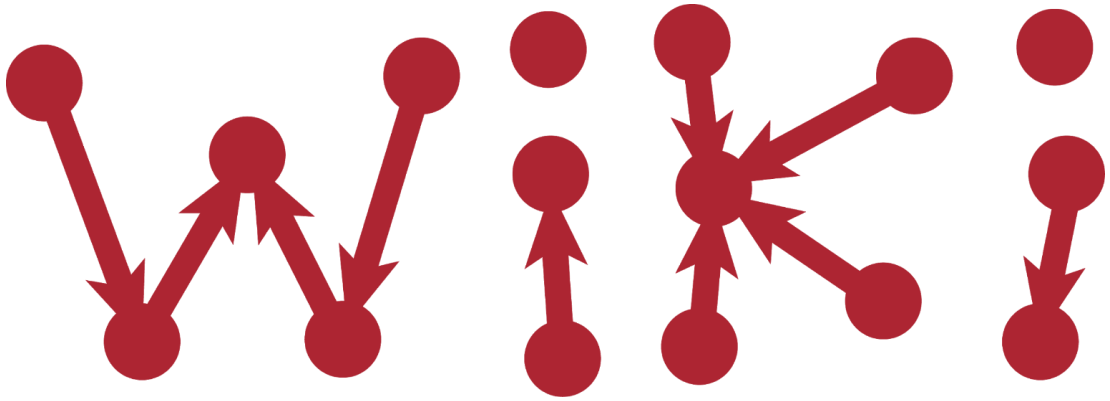


Wikipedia Analysis Using a Keyword Based Graph

Project – A Network Tour of Data Science

Marc GLETTIG
Matthias MINDER
Yves RYCHENER
Charles TROTIN
Ecole Polytechnique Fédérale de Lausanne

Problem Definition

Network science

From Wikipedia, the free encyclopedia

For other uses, see Network (disambiguation).

Network science is an academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors represented by *nodes* (or *vertices*) and the connections between the elements or actors as *links* (or *edges*). The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, and social structure from sociology. The United States National Research Council defines network science as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena."[1]

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Can a keyword based graph predict the hyperlink network of Wikipedia?

Contents

- Graph construction using text mining
- Keyword based graph analysis and comparison
- New links suggestions

The Data

- Subsampled Wikipedia data, 4’604
- Remove isolated nodes
- Remove term definition sites

Result:
4’549 articles, 118’809 edges,
average degree: 26.1

Dark Ages

From Wikipedia, the free encyclopedia

Dark Ages or **Dark Age** may refer to:

History and sociology [edit]

- European Early Middle Ages, often referred to as the *Dark Ages*, or the European Middle Ages in general (5th to 15th centuries AD), particularly:
  - Migration Period of c. 400 to 800 AD
  - *Saeculum obscurum* or "dark age" in the history of the papacy, running from 904 to 964 AD
- Dark Ages (historiography), the use of the term *Dark Ages* by historians and lay people

Text Mining

1. Text cleaning + Bag of words representation
2. Term Frequency Calculation
3. Inverse document frequency

$$IDF(w) = 1 + log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ w})$$

4. Combining TF and IDF

Graph Construction

Inference Methods:

- K Nearest Neighbors. Directed
- Calculate pairwise Euclidean distance, retain closest as edge. Undirected
- (Calculate Cosine similarity, retain most similar. Undirected)

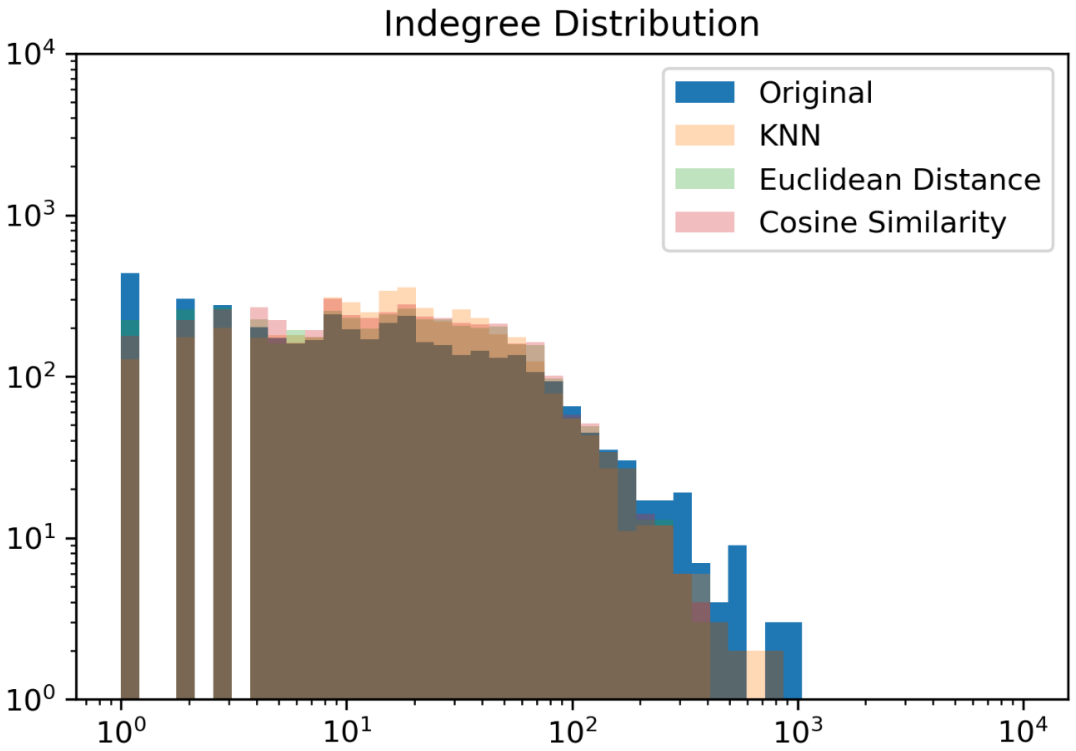
Tuned to have the same amount of edges

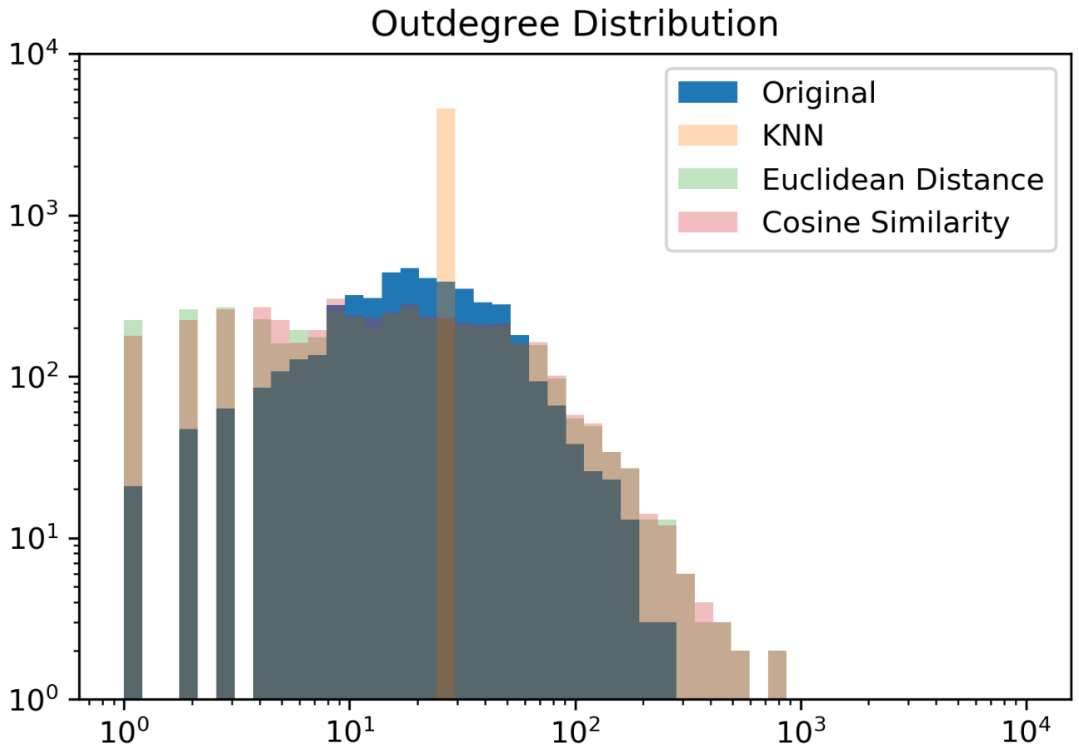
Results: Edge Similarity

Measure percentage of identical edges between constructed and original network

- K Nearest Neighbors: 23.92%
- Closest Distance: 23.60%
- (Cosine Similarity: 22.84%)

Results: Degree Distribution





PageRank Algorithm

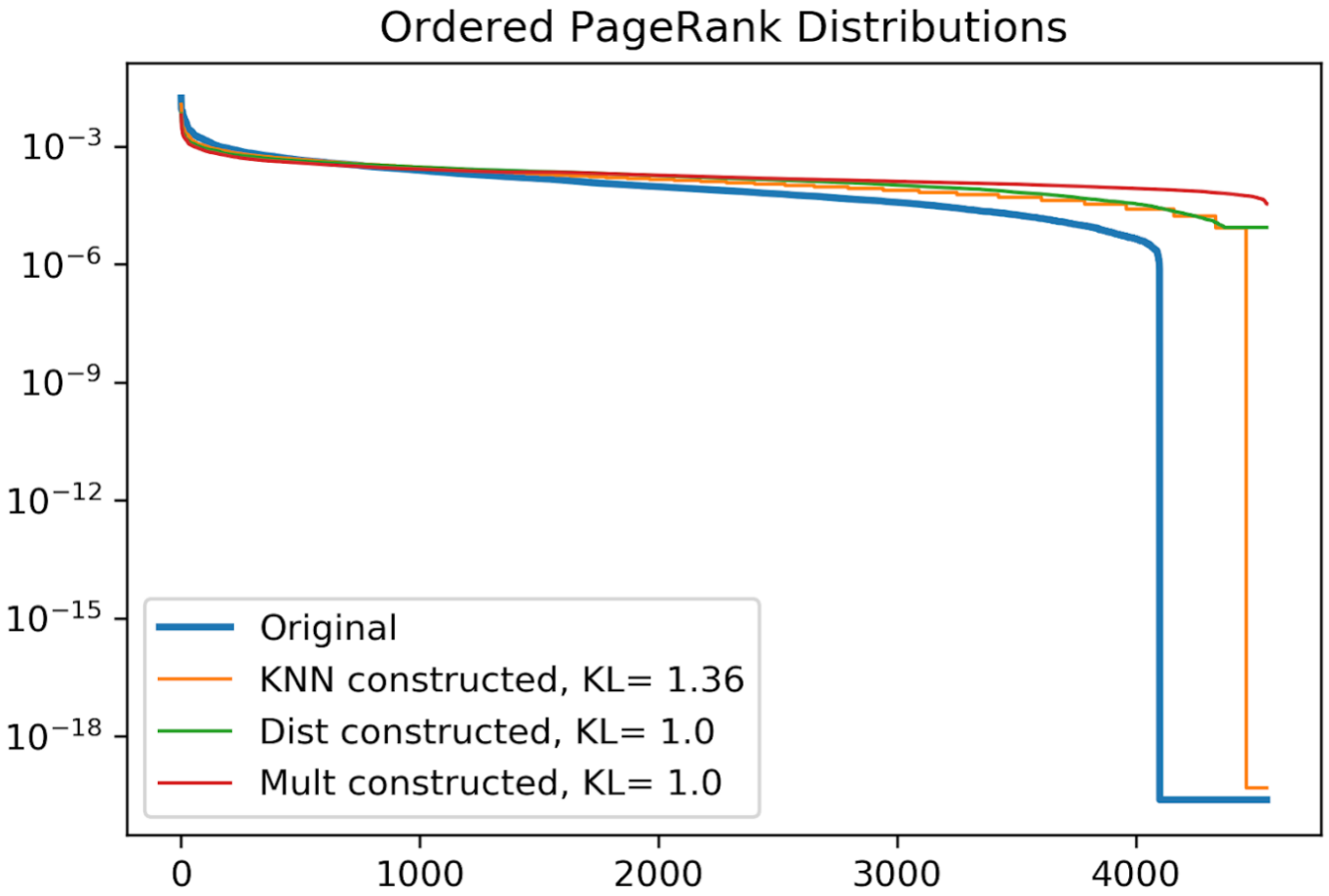
1 View Network as Markov Chain navigated by surfer

2 Calculate stationary distribution of surfer

Most Important Articles according to PageRank

Original	KNN	Distance	Multiplication
United States	United States	United States	United States
United Kingdom	Modern history	Modern history	Modern_history
Scientific classification	United Kingdom	Bird	Bird
Europe	England	United Kingdom	United Kingdom
England	20th Century	New York City	New York City

Results: PageRank



Ordered PageRank probabilities and KL-Divergence

Link Suggestions

Steps:

- Calculate pairwise euclidean distance between articles
- Identify edges present in KNN-Network and not in original network
- Sort the edge distances, retain on average three edges per node

Why KNN?

- Ensures that a maximum of 26 links are suggested for each article
- Don’t only suggest links for central nodes

Measuring prediction quality

- Manual quality assessment of suggested links

Suggestion Examples

Observations:

- Links from articles to more general category
- Still needs human guidance

Five Suggested Links Selected at Random

Number	From	To	
1	Avacha Volcano	Galeras	
2	Byzantine Empire	6th century	✓
3	A Tale of a Tub	Augustan literature	✓
4	Post-glacial rebound	Sea level rise	✓
5	Lake Chad	Lake Superior	

Conclusion & Suggestions

Conclusions

- Degree distribution similar, but asymmetry not captured
- Other factors influence hyperlink network
- Reasonably good link suggestion

Further ideas

- More sophisticated directed network construction, such as dynamic k in kNN
- Group of words in text mining, use word embeddings
- Article metadata inclusion

Question

From Wikipedia, the free encyclopedia

To ask questions about Wikipedia, see Wikipedia:Questions.
For other uses, see Question (disambiguation).

A **question** is an utterance which typically functions as a request for information. Questions can thus be understood as a kind of illocutionary act in the field of pragmatics or as special kinds of propositions in frameworks of formal semantics such as alternative semantics or inquisitive semantics. The information requested is expected to be provided in the form of an answer. Questions are often conflated with interrogatives, which are the grammatical forms typically used to achieve them. Rhetorical questions, for example, are interrogative in form but may not be considered true questions as they are not expected to be answered. Conversely, non-interrogative grammatical structures may be considered questions as in the case of the imperative sentence "tell me your name".