Improving Search and Retrieval in Digital Libraries by Leveraging Keyphrase Extraction Systems

Wei Jin
Corina Florescu

University of North Texas

Joint Conference on Digital Libraries, 2018
This tutorial presents:

- the task of keyphrase extraction i.e., state-of-the-art approaches for extracting keyphrases from various types of documents; benefits and limitations of these approaches;

- keyphrase extraction in context of digital libraries i.e., several digital library applications that leverage keyphrase extraction; An analyze of keyphrase extraction systems to understand which better suits the problem of digital libraries;

After completing this tutorial, you will know:

- the main supervised and unsupervised approaches for keyphrase extraction

- their advantages and disadvantages and how to choose the one that best suit your task

- how to use them in order to improve various tasks in digital libraries;
Corina Florescu

Affiliation: University of North Texas

Several Publications:
Corina Florescu, Wei Jin. Learning Feature Representations for Keyphrase Extraction. In: AAAI 2018

Corina Florescu, Cornelia Caragea. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. ACL 2017


A Markov Chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov Chain is a type of Markov process that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of Markov processes [...]. Markov Chains have many applications as statistical models of real-world processes, [...]. The algorithm known as PageRank, which was originally proposed for the Internet search engine Google, is based on a Markov process.
A **Markov chain** is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A **Markov chain** is a type of Markov process that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of Markov processes[...]. **Markov chains** have many applications as statistical models of real-world processes, [...]. The algorithm known as , which was originally proposed for the Internet search engine Google, is based on a Markov process.
A **Markov chain** is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A **Markov chain** is a type of Markov process that has either discrete state space or discrete index set, [...] Random walks on integers and the gambler’s ruin problem are examples of Markov processes[...]. **Markov chains** have many applications as statistical models of real-world processes, [...]. The algorithm known as **PageRank**, which was originally proposed for the Internet search engine Google, is based on a Markov process.
A Markov Chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov Chain is a type of Markov process that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of Markov processes [...]. Markov Chains have many applications as statistical models of real-world processes, [...]. The algorithm known as PageRank, which was originally proposed for the Internet search engine Google, is based on a Markov process.

- **Potential Keyphrases:**
  - Markov chain, Markov process, stochastic process
Why Keyphrase Extraction?

- Enable the reader to quickly determine whether a piece of text (e.g., news article, web page) is in the reader’s fields of interest.
- When they are used by a search engine, the goal is to make the search more precise.
Many documents do not have associated keyphrases.

Given a large number of text documents existing today, human labeling is impractical.

Automatic techniques are required for extracting keyphrases from various documents.

Automatic approaches to keyphrase extraction have been developed along two lines of research: supervised and unsupervised.
Automatic Keyphrase Extraction

Identify Candidate Phrases
- text units from the target document (e.g., single or multi-word phrases)

Supervised
- represent each candidate phrase with a set of features
- train a machine learning algorithm to predict keyphrases

Unsupervised
- compute a score for each candidate phrase

Output
- output keyphrases using classifier’s confidence scores
  e.g., all phrases with confidence score >50%

Output
- rank keyphrases based on their scores
- output the first k ranked phrases as keyphrases
A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov chain is a type of Markov process that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of Markov processes[...]. Markov chains have many applications as statistical models of real-world processes, [...]. The algorithm known as PageRank, which was originally proposed for the Internet search engine Google, is based on a Markov process.
A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov chain is a type of Markov process that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of Markov processes[...]. Markov chains have many applications as statistical models of real-world processes, [...]. The algorithm known as PageRank, which was originally proposed for the Internet search engine Google, is based on a Markov process.

- **N-grams:** A; A Markov; A Markov chain; A Markov chain is; Markov; Markov chain; Markov chain is; Markov chain is a;
A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property. A Markov chain is a type of Markov process that has either discrete state space or discrete index set, which means that the time is represented by a discrete variable. This property(ref) satisfies the Markov property.

A Markov chain is a type of Markov process that has either discrete state space or discrete index set, which means that the time is represented by a discrete variable. This property(ref) satisfies the Markov property.

Wei Jin Corina Florescu (University of North Texas) Improving Search and Retrieval in Digital Libraries by Leveraging Keyphrase Extraction Systems Joint Conference on Digital Libraries, 2018
A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

In probability theory and related fields, a Markov process, named after the Russian mathematician Andrey Markov, is a stochastic process that satisfies the Markov property.

A Markov chain is a type of Markov process that has either a discrete state space or a discrete index set.
Supervised Approaches

- Supervised approaches for keyphrase extraction aim to train a machine learning algorithm to determine the keyphrases of a document [Hulth(2003), Medelyan et al.(2009)Medelyan, Frank, and Witten]

- Treat a document as a set of phrases, which have to be classified as either positive or negative examples.

- This is a classical machine learning problem of learning from examples which conveys that we need the following components:
  - items to be classified (phrases)
  - features to represent a phrase
  - true labels (“correct” keyphrases)
Feature Design

- Find some properties/characteristics of these phrases that distinguish keyphrases from non-keyphrases:

Text Selections from Wikipedia

A **Markov chain** is a **stochastic model** describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In probability theory and related fields, a **Markov process**, named after the Russian mathematician Andrey Markov, is a **stochastic process** that satisfies the Markov property. A **Markov chain** is a type of **Markov process** that has either discrete state space or discrete index set, [...]. Random walks on integers and the gambler’s ruin problem are examples of **Markov processes** [...]. **Markov chains** have many applications as statistical models of real-world processes, [...]. The algorithm known as PageRank, which was originally proposed for the Internet search engine Google, is based on a **Markov process**.
Feature Design

- Find some properties/characteristics of these phrases that distinguish keyphrases from non-keyphrases:
  - The first position of a phrase in the document
    [Frank et al.(1999)]
  - The frequency of a phrase in the document [Hulth(2003)]
  - The number of words per phrase
    [Medelyan et al.(2009)]
  - Wikipedia statistics (e.g., the phrase ia a link in Wikipedia or the phrase is part of a Wikipedia title page)
    [Medelyan et al.(2009)]
  - The location of a phrase in different sections of a document
    [Nguyen and Luong(2010)]
  - The existence of a phrase in the citation contexts
    [Caragea et al.(2014)]
In the training process, we have to provide the “true” class of items to be classified.

To obtain a set of “correct” keyphrases for each document in the training collection, we rely on human labeling/annotation.
What other people say ...

- mad cow disease
- british beef exports
- bovine spongiform encephalopathy
- unilateral ban
- british exports
- restrictions
- bse affected cows
- Germany
“Correct Keyphrases” of Documents

Problems with “correct” keyphrases:

- the task is inherently subjective, i.e., keyphrases assigned by one annotator are not the only correct ones
  [Sterckx et al.(2016)Sterckx, Demeester, Develder, and Caragea]

- The agreement between annotators is usually very low

- Semantically equivalent keyphrases are being annotated in different forms, e.g., “Louis Michel” vs. “Prime Minister Louis Michel”, “bovine spongiform encephalopathy” vs “bse”

- Human annotations may be redundant, e.g., “british beef exports” vs. “beef exports” vs. “beef exports”

- **These have consequences for training, developing and evaluating supervised models**
In the unsupervised line of research, keyphrase extraction is formulated as a ranking problem where each phrase receives a score based on various measures:

- TF-IDF
- graph-based ranking methods (e.g., PageRank) [Mihalcea and Tarau (2004), Liu et al. (2010)]
- similarity scores Bennani-Smires et al. (2018)

Phrases are ranked based on their scores and top $k$ phrases are retrieved as keyphrases of that document.
Let $D$ be a collection of documents such that $|D| = N$ and $t$ a term.

$$tf-idf(t) = tf(t) \cdot idf(t) = tf(t) \cdot \log\frac{N}{|d \in D : t \in d| + 1}$$

**Intuition:**
- $tf$ - the more often a term occurs in a document, the more representative it is of this document
- $idf$ - the more documents contain a term, the least discriminating it becomes

The ranking based on $tf-idf$ has been shown to work well in practice, despite its simplicity.

**Is tf-idf domain-independent?**
PageRank is a link analysis algorithm and it assigns a numerical value to each document, with the purpose of "measuring" its relative importance within the set.

The underlying assumption is that more important websites are likely to receive more links from other websites.
Window = 2

Markov chain is a type of Markov process that has either discrete state space or discrete index set, [...]
Many graph-based extensions have been proposed, which aim at modeling various types of information:

- a local neighborhood of the target document corresponding to its textually-similar documents [Wan and Xiao(2008)]
- information from Wikipedia or WordNet Martinez-Romo et al.(2016)
- information from the citation network [Gollapalli and Caragea(2014)]
- the position information of words [Florescu and Caragea(2017)]
- topic models e.g., LDA [Liu et al.(2010)Liu, Huang, Zheng, and Sun]
Summary on Keyphrase Extraction

- **Supervised Approaches**
  - Require labeled training data
  - Allow for more expressive feature design
  - Are (commonly) domain-dependent

- **Unsupervised Approaches**
  - DO NOT require labeled training data
  - Are (usually) domain-independent
  - Less flexible than supervised models
## Case Study

### Supervised (Maui)
- mad cow disease
- bovine spongiform encephalopathy
- British agriculture
- UK
- BSE
- scientific evidence
- ban
- Germany
- single market
- beef exports

### What people say ...
- mad cow disease
- british beef exports
- bovine spongiform encephalopathy
- unilateral ban
- british exports
- restrictions
- bse affected cows
- Germany
Case Study

Unsupervised (TopicRank)
- Germany
- European Commission
- British beef export
- Ban
- Legal action
- BSE
- Mad cow disease
- Human
- Live cattle
- UK beef sector

What other people say ...
- Mad cow disease
- British beef exports
- Bovine spongiform encephalopathy
- Unilateral ban
- British exports
- Restrictions
- BSE affected cows
- Germany
Case Study

Supervised (Maui)
- mad cow disease
- bovine spongiform encephalopathy
- british agriculture
- UK
- BSE
- scientific evidence
- ban
- Germany
- single market
- beef exports

Unsupervised (TopicRank)
- Germany
- European Commission
- British beef export
- ban
- legal action
- BSE
- mad cow disease
- human
- live cattle
- UK beef sector
Final Notes on Keyphrase Extraction

- The difficulty of the task increases with the length of the input document.

- It is easier to extract keyphrases from structured documents.

- Some documents present topic change (e.g., meeting transcripts, chats); in a conversation, the topics change as the interaction moves forward in time; **topic change detection is not always easy**.

- Topic correlation - keyphrases of a document are typically related to each other in research papers or news articles; this observation does not necessarily hold for informal text, where people can talk about any number of potentially uncorrelated topics.
Final Notes on Keyphrase Extraction

- Overgeneration errors (e.g., third world, the third world economy, third world nation, third world country); Possible Solution: background knowledge extracted from external databases, clustering.

- Infrequency errors Possible Solution: We need to find a way to boost its importance (frequency), e.g., using the frequency of its related counterparts

- Redundancy errors; Possible Solution: background knowledge extracted from external databases, clustering.

- Evaluation errors; Possible Solution: one possibility is to conduct human evaluations

(*) For more details check the following survey [Hasan and Ng(2014)]
Digital Libraries

- What is a digital library?
  - Google Scholar, CiteSeer, Internet Archive, Oxford Text Archive

- What items should be in a digital library?

- How can the items be organized to support knowledge discovery?
What is a digital library?

- online database of digital objects (digitized or born-digital)
- tools for organizing, searching, and retrieving content from the collection
- those tools should be personalized to support user needs
- Google Scholar, CiteSeer, Internet Archive, Oxford Text Archive

What items should be in a digital library?

How can the items be organized to support knowledge discovery?
What is a digital library?
- online database of digital objects (digitized or born-digital)
- tools for organizing, searching, and retrieving content from the collection
- those tools should be personalized to support user needs

Google Scholar, CiteSeer, Internet Archive, Oxford Text Archive

What items should be in a digital library?
- we usually store text documents
- photographs, videos, sensor data

How can the items be organized to support knowledge discovery?
What is a digital library?
- online database of digital objects (digitized or born-digital)
- tools for organizing, searching, and retrieving content from the collection
- those tools should be personalized to support user needs
- Google Scholar, CiteSeer, Internet Archive, Oxford Text Archive

What items should be in a digital library?
- we usually store text documents
- photographs, videos, sensor data

How can the items be organized to support knowledge discovery?
- depending on the task, you may want to:
  - cluster/group the items based on various criteria
  - classify the items in different categories
  - consider a graph representation of items
How can we use keyphrase extraction systems to support digital libraries (searching, retrieval, knowledge discovery)?

- indexing
- classification
- clustering
- summarization
- hyperlink browsing
Indexing - the process of describing the content of a document by a set of terms (words or phrases that captures the essence/idea of the document)

- In digital libraries, indexing is usually performed by professional indexers

- Indexing is basically performed in a 2-step process: (1) identify terms/concepts that describe the document; (2) map those terms to a controlled vocabulary
**Indexing** - the process of describing the content of a document by a set of terms (words or phrases that captures the essence/idea of the document)

- In digital libraries, indexing is usually performed by professional indexers

- Indexing is basically performed in a 2-step process: (1) identify terms/concepts that describe the document; (2) map those terms to a controlled vocabulary
Indexing - the process of describing the content of a document by a set of terms (words or phrases that captures the essence/idea of the document)

- In digital libraries, indexing is usually performed by professional indexers

- Indexing is basically performed in a 2-step process: (1) identify terms/concepts that describe the document; (2) map those terms to a controlled vocabulary

- Automatic index term assignment is a great improvement for digital libraries
Indexing

Things to consider when using the keyphrase extraction systems for automatic term indexing:

- Characteristics of text (documents) being stored
  - are the documents that I want to store from the same domain?
  - do the documents have a particular structure?
  - how long are the documents?
  - what language the documents are written in?

- Specificity of terms to be assigned
  - more general terms, e.g., *machine learning*;
  - more specific terms, e.g., *Naive Bayes*

People may want to learn about a collection, what it contains, and how well it covers a particular topic;

Although most systems provide a brief description of the collections contents (e.g. computer science technical reports), they rarely display the range of topics covered.

We can use keyphrase extraction systems to provide information about the top-level contents of a document collection.

We may want to cluster documents into topics and then use keyphrase extraction to assign labels to each cluster/topic (KE from topic-related documents has shown to work well)

The core idea to access information in WWW is to navigate between documents (web pages) via embedded hyperlinks.

Wikipedia provides users with such a feature.

Many digital libraries do not contain browsable links.

We can leverage keyphrase extraction systems to support “hyperlink navigation.”
We can use keyphrase extraction to insert a link anchor into the text whenever a phrase occurs that is a keyphrase in other documents.

When the user selects a phrase, a new frame (window) is generated that lists the documents for which that phrase is a keyphrase;

Selecting a document from the list loads it. You may also provide a short summary of the document when the mouse is over its title (KE can be employed to generate the summary).

Keyphrase-based Recommender System

- Compute a user profile based on the available information, e.g., documents the user reads or clicks, fields of interest found in the user profile
- Extract keyphrases from papers which are relevant to a specific user
- Then, in order to compute the relevance of a new article, the user profile is compared with the keyphrase list extracted from that article

Can you think of some other application of keyphrase extraction in digital libraries?

Tell me about some challenges that you faced while working with digital libraries

Can you tell me some future directions in digital libraries?
Future Directions in Keyphrase Extraction

- Use feature learning or representation learning to automatically discover characteristics that explain some structure underlying the data (i.e., patterns that distinguish keyphrases from non-keyphrases).
- Use more powerful models (e.g., neural networks) which can consider the sequential nature of text when identifying keyphrases.
- Consider the extent to which a keyphrase represents the content of a document.
Questions?

• ???

• For further questions you can email me at: CorinaFlorescu@my.unt.edu
You can find my research at: https://corinaflorescu.github.io/cs/
Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2014.
Representing topics labels for exploring digital libraries.

Opinion expression mining by exploiting keyphrase extraction.
In *Asian Federation of Natural Language Processing*.

Finding topic trends in digital libraries.
Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014.
Citation-enhanced keyphrase extraction from research papers: A supervised approach.

Felice Ferrara, Nirmala Pudota, and Carlo Tasso. 2011.
A keyphrase-based paper recommender system.

Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents.


References IV

Improving browsing in digital libraries with keyphrase indexes.

Corephrase: Keyphrase extraction for document clustering.
In *Machine Learning and Data Mining in Pattern Recognition*, pages 265–274. Springer.

Kazi Saidul Hasan and Vincent Ng. 2014.
Automatic keyphrase extraction: A survey of the state of the art.
Improved automatic keyword extraction given more linguistic knowledge.

Steve Jones and Gordon Paynter. 1999.
Topic-based browsing within a digital library using keyphrases.
In *Proceedings of the fourth ACM conference on Digital libraries*, pages 114–121. ACM.

Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications.


Olena Medelyan and Ian H Witten. 2006.
Thesaurus based automatic keyphrase indexing.

Textrank: Bringing order into text.

Thuy Dung Nguyen and Minh-Thang Luong. 2010.
Wingnus: Keyphrase extraction utilizing document logical structure.


Single document keyphrase extraction using neighborhood knowledge.
8:855–860.

Hongyuan Zha. 2002.
Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering.
pages 113–120.