# A Visual Approach for Interactive Keyterm-Based Clustering (Nourashrafeddin et al., 2018)

## Paper Analysis

Eric Macedo Cabral

Universidade de São Paulo (USP)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Laboratório de Visualização, Imagens e Computação Gráfica (VICG)

*cabral.eric@usp.br*

September 17, 2018

# Abstract


- Keyterm-based approach is arguably intuitive for users to direct text-clustering processes and adapt results to various applications in text analysis.
- This article first presents a text-clustering algorithm that can easily be extended into an interactive algorithm.
- Visualizations are provided for the whole collection as well as for detailed views of document and cluster relationships.



CCS Concepts:
• **Information systems** → **Clustering**;
• **Human-centered computing** → Visual analytics;

**Keywords:** document clustering, keyterm-based clustering, visualization, interactive.

# Table of Contents

# Introduction I

- By grouping similar documents, clustering algorithms provide precious information about topics in text collections.
  - Traditionally, no user-effort is targeted and the user has minimum interaction with the clustering process.
  - By putting the user in the clustering process, the results are more likely to satisfy his needs.

# Introduction II

Three main categories of interaction in text clustering:

- **Document supervision:** It relies on training documents to coordinate the clustering process (pre-labeled), seeds.
- **Keyterm supervision:** The main tasks of this approach are extracting relevant keyterms from documents and clusters, presenting them to the user, collecting feedback, and incorporating them in the clustering process.
- **Hybrid keyterm-documetn supervision:** It uses both document and keyterm supervision

# Introduction III

- Keyterm supervision is more intuitive and requires less user effort.
  - It reflects to the individual's preferences.
- Clustering documents by groups of terms help the user to target the sense of the document by the context of the terms, not a single term.

# Table of Contents

# Related Work

## Visual Text Clustering Using High-Resolution Display

- All the works in this section require the user to read the documents and label them. Too much user effort.
- Working based on relevant terms, on the other hand, allows us to express topics of documents in a more abstract form.
  A few keyterms can represent hundreds of relevant documents

# Related Work

## Visual Text Clustering Using Semi-Supervision

- All the works in this section require the user to move the instances manually or to declare document seeds. Too much user effort.
- Examining all data objects and moving them on a 2D plot is time-consuming, especially for large collections
- Also, a single document does not contain enough information to effectively build a cluster around it [Aggarwal and Zhai (2013)].

# Related Work

## Visual Text Clustering Based on Topic Modeling

- Latent Dirichlet Allocation (LDA).
  - Generate initial topics.
- Requires the user to define weights to the generated topics.
  - Defining a ranked list of terms is more intuitive.
- Hard clustering.
  - Soft clustering is more informative for the user.

# Table of Contents

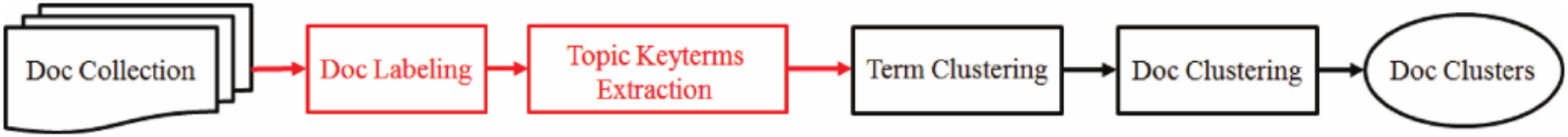
# Lexical Double Clusterer (LDC) I

- The idea behind the algorithm is that before finding document clusters it is better to focus on term clusters and the keyterms that represent topics.
- It uses a matrix of documents-terms, obtained with the Bag of Words (BoW) Model.
- Term frequency-inverse document frequency (TF-IDF) is used as feature values to indicate the importance of terms in documents.
  - Discriminative terms hava a higher TF-IDF in a subset of documents.
- The term clusters are obtained with the fuzzy c-means algorithm. Which means they are soft clusters.
- By finding the terms clusters, it's possible to find good document seeds that well represent that topic space (Hybrid supervision).

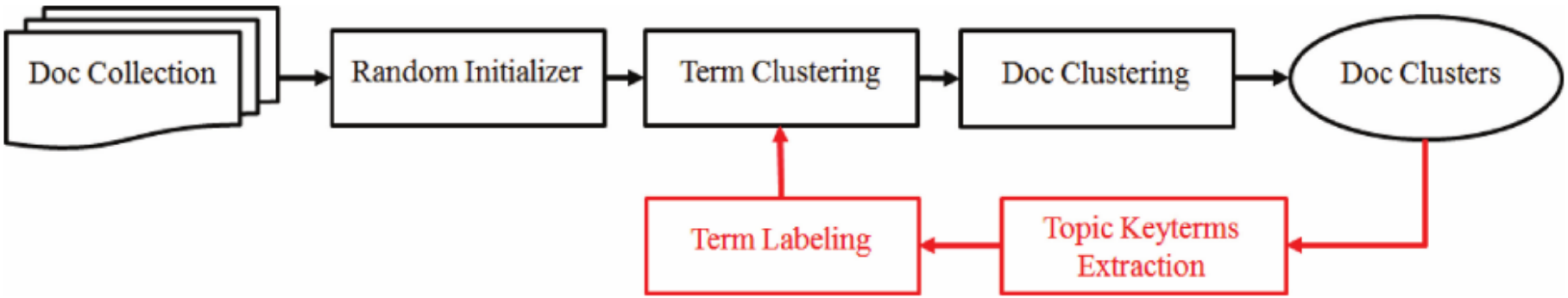
# Lexical Double Clusterer (LDC) II

LDC can be easily adapted for interactive use. In a way that we can categorize interactions by:

- **Document-supervised:** Documents seeds.




- **Term-surpervised:** Terms manipulations (add, remove, move...).



# Lexical Double Clusterer (LDC) III

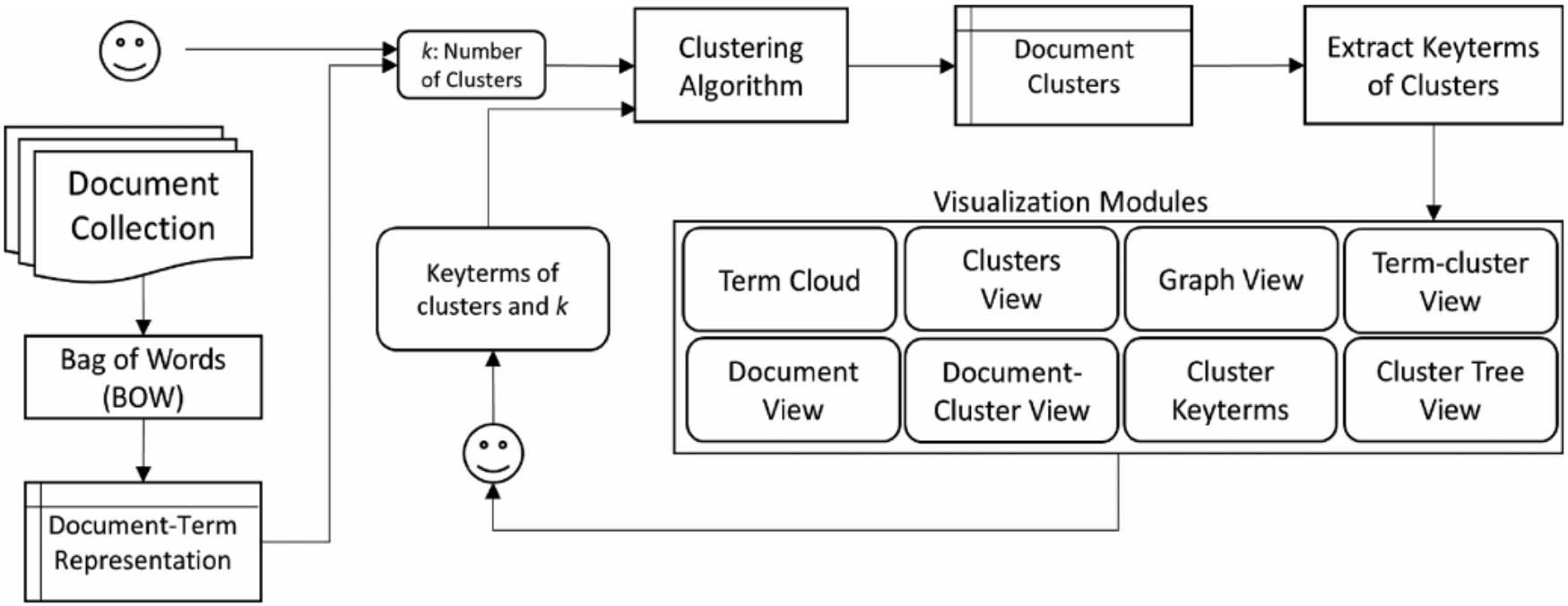
Datasets used in the tests:

| Dataset Name | No. of Classes | No. of Documents | Description |
|---|---|---|---|
| 20newsgroups | 20 | 18,821 | All news articles in 20newsgroups after removing duplicates |
| NewsGroup9 | 9 | 720 | Subset of news articles of 20newsgroups dataset |
| NewsRelated | 3 | 2,624 | Subset of news articles of 20newsgroups dataset |
| Reuters8 | 8 | 7,674 | Articles of Reuters-21578 in eight categories |
| Classic-4 | 4 | 7,095 | Scientific abstracts in four categories |
| WebKB | 4 | 4,168 | Webpages of computer science departments |
| SMS | 2 | 5,479 | A public set of text messages |
| BBC Sport | 5 | 737 | BBC Sport news articles collected in 2004-2005 |
| News 2006 | — | 1,747 | News feeds collected in 2006 from Associated Press, CNN, and Reuters |
| NewsSeparate | 13 | 381 | A subset of News 2006 dataset categorized into 13 classes manually |

# Table of Contents

# VIS-KT

## A visual Framework to Support Keyterm-based Clustering I



# VIS-KT

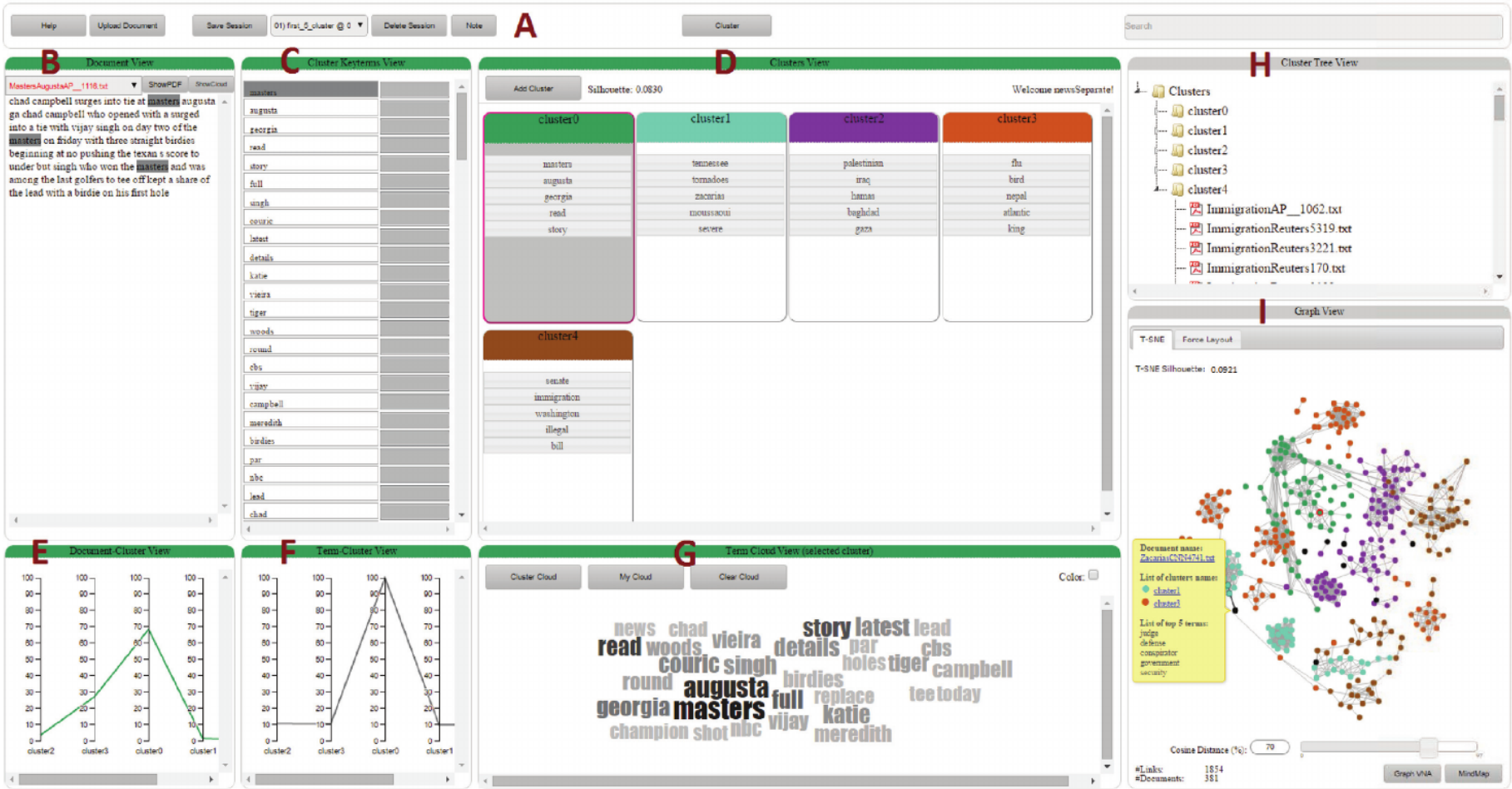
## A visual Framework to Support Keyterm-based Clustering II

Main visual components:

- Document View
- Cluster Keyterm View
- Cluster View
- Term Cloud View
- Cluster Tree View
- Graph View

# VIS-KT

## A visual Framework to Support Keyterm-based Clustering III



Further tests have shown improvement as the user interacted with the data:

*NewsSeparate* in Each Iteration

| | Average Silhouette | | | |
|---|---|---|---|---|
| | Iteration1 | Iteration2 | Iteration3 | Iteration4 |
| Clustering | 0.0830 | 0.1718 | 0.2011 | 0.2224 |
| t-SNE | 0.0522 | 0.2305 | 0.3530 | 0.3385 |
| Force Layout | 0.0474 | 0.2423 | 0.3774 | 0.4681 |

The silhouette indices of t-SNE and force layout are independent of the user interaction. The silhouette index of the clustering shows the quality of clustering.

# Table of Contents

# Conclusion

## Contributions

- Interactive text-clustering visualization with keyterm labeling support.
- LDC.

## Positive Points

- The method proposed presents good results.
- The tests and the results were expressive and informative.
- It is open-source[a].

[a] https://github.com/ehsansherkat/IDC/

## Negative Points

- None that I could find.

# Questions?

# References I

S. Nourashrafeddin, E. Sherkat, R. Minghim, and E. E. Milios, "A Visual Approach for Interactive Keyterm-Based Clustering," *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 1, pp. 1–35, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3185338.3181669

C. C. Aggarwal and C. X. Zhai, *Mining text data*, 2013, vol. 9781461432.