

Text Summarization for Compressed Inverted Indexes and Snippets

Advisor : Dr. Chris Pollett

Committee members : Dr. Sami Khuri, Dr. Ronald Mak

Presented by
Mangesh Dahale

Agenda

- Introduction
- Preliminary Work
- Summarization in Yioop
- Experiments
- Conclusion
- Demo

Introduction

- Search engines are often the first source of information when we want to do any research
- Text summarization is a technique to generate a concise summary of a larger text.
- The major challenge in summarization
 - distinguishing the more informative parts of a document from the less informative parts.

Text Summarization

- Text summarization is a three step process :
 1. Selection of salient portions of text
 2. Aggregation and abstraction of this information
 3. Presentation of the final summary text
- Two main approaches:
 - Extraction based
 - Abstraction based
- Use of summaries for:
 - Indexing
 - Query results

Preliminary Work

- Three summarization techniques
 1. Intersection method
 2. Centroid method
 3. TF-ISF method.
- Evaluate the performance of these three methods to find the best summarization method.

Text Summarization using Intersection Function

- Based on the TextRank algorithm by Mihalca R. and Tarau P.
- Represent the complete text as a graph.
 - Vertices - sentences from the text
 - Weighted Edges - similarity score of two sentences
- Sentence dictionary
 - Sentence and Total score of each sentence

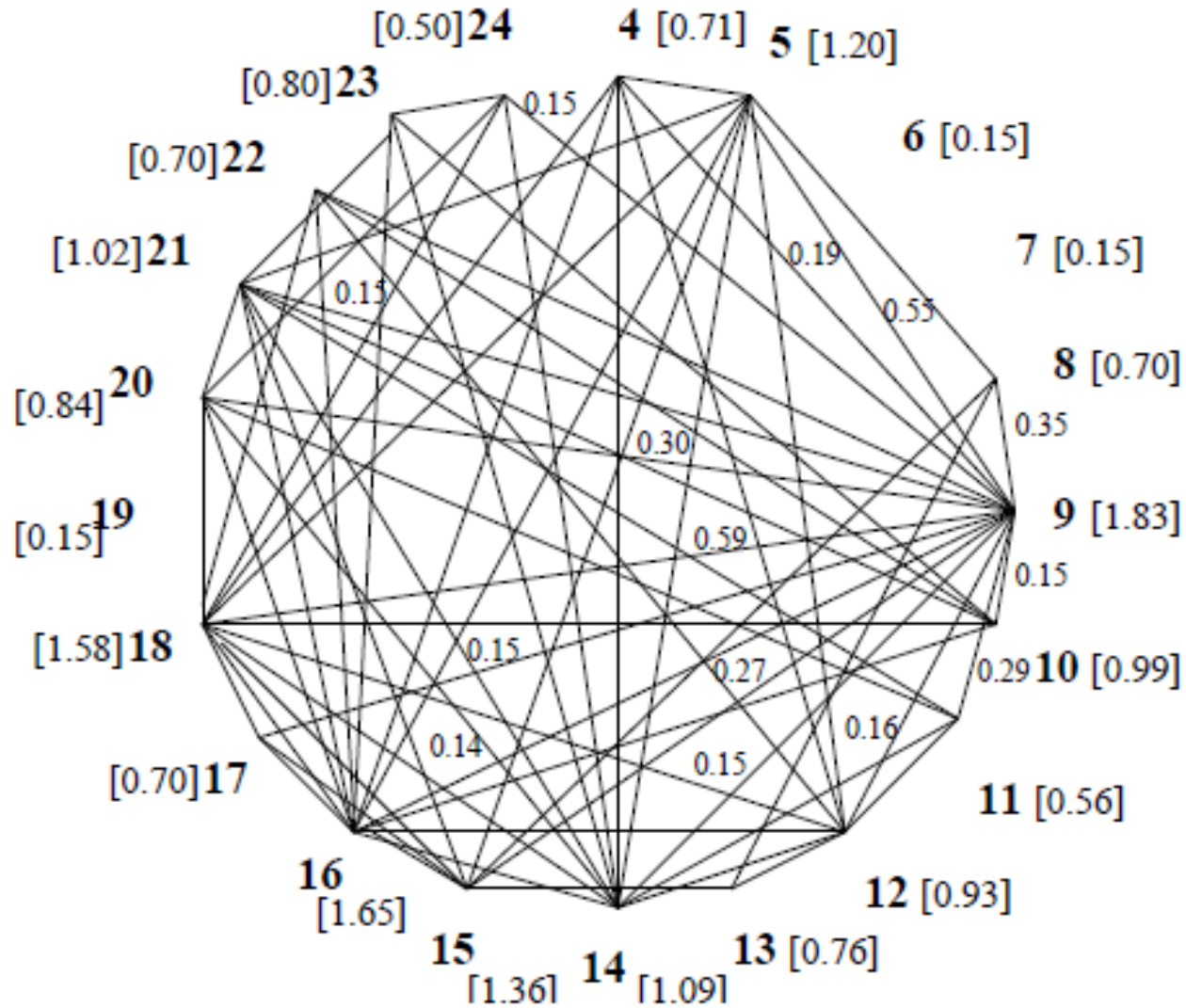
Text Summarization using Intersection Function

- Intersection Function:

$$f(s_1, s_2) = \frac{|\{w \mid w \text{ in } s_1 \text{ and } w \text{ in } s_2\}|}{(|s_1| + |s_2|) / 2}$$

- If two sentences have a good intersection, they probably hold the same information.
- If one sentence has a good intersection with many other sentences, it probably holds some information from each one of them.

Sample graph build using TextRank algorithm



Text Summarization using centroid algorithm

- Multi-document summarizer
- What is a centroid?
 - A centroid is a set of words that are statistically important to a cluster of documents
- Each document is represented as a weighted vector of TF-IDF

Centroid Algorithm

- Generate a centroid by using only the first document in the cluster.
- As new documents are processed, their TF-IDF values are compared with the centroid using the formula

$$sim(D, C) = \frac{\sum_k (d_k \cdot c_k \cdot idf(k))}{\sqrt{\sum_k (d_k)^2} \sqrt{\sum_k (c_k)^2}}$$

Text Summarization using TF-ISF

- Represent the document as the set of sentences
- $S = \{s_1, s_2, \dots, s_n\}$ represents all sentences in document
- $T = \{t_1, t_2, \dots, t_m\}$ represents all the terms in S
- w_{ij} associated with term t_j in sentence s_i is calculated by the scheme tf–isf.

$$w_{ik} = tf_{ik} \cdot \log(n / n_k)$$

The Cosine Similarity

$$\text{sim}(s_i, s_j) = \frac{\sum_{k=1}^m w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^m w_{ik}^2 \cdot \sum_{k=1}^m w_{jk}^2}}, \quad i, j = 1, \dots, n$$

- ▶ Coverage

Coverage means that the generated summary should cover all subtopics as much as possible

- ▶ Diversity

Sentences in a summary should have little overlap with one another in order to increase diversity

Coverage and Diversity

- ▶ Coverage

$$f_{\text{cover}}(X) = \text{sim}(O, O^S) + \text{sim}(O, s_i)$$

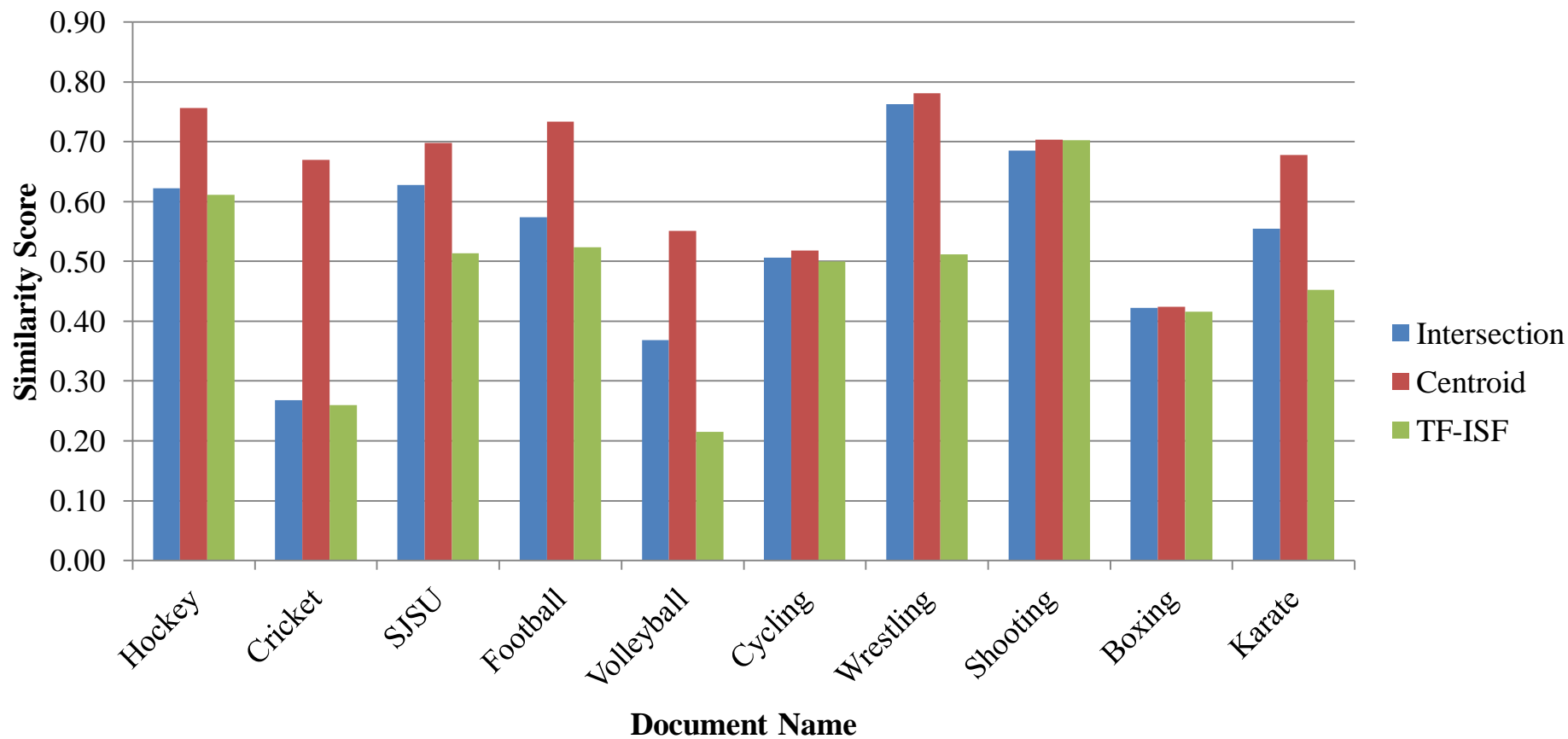
- ▶ Diversity

$$f_{\text{diver}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (1 - \text{sim}(s_i, s_j)) x_i x_j$$

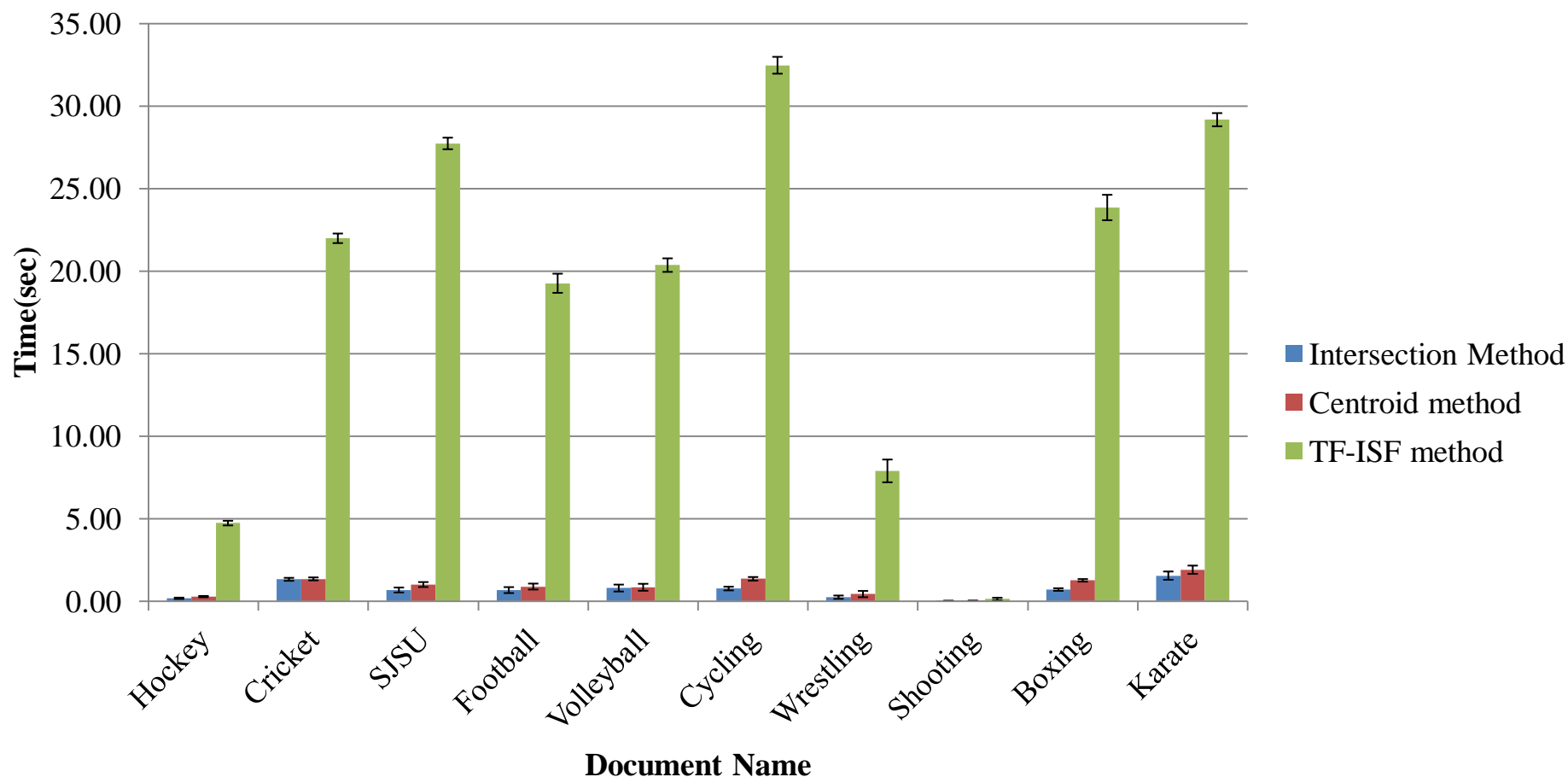
Evaluation of three methods

- Made a set of ten documents and their human generated summary.
- Ran all three methods on same set of documents.
- Calculated the cosine similarity between the human generated summary and the summary generated by all three methods.

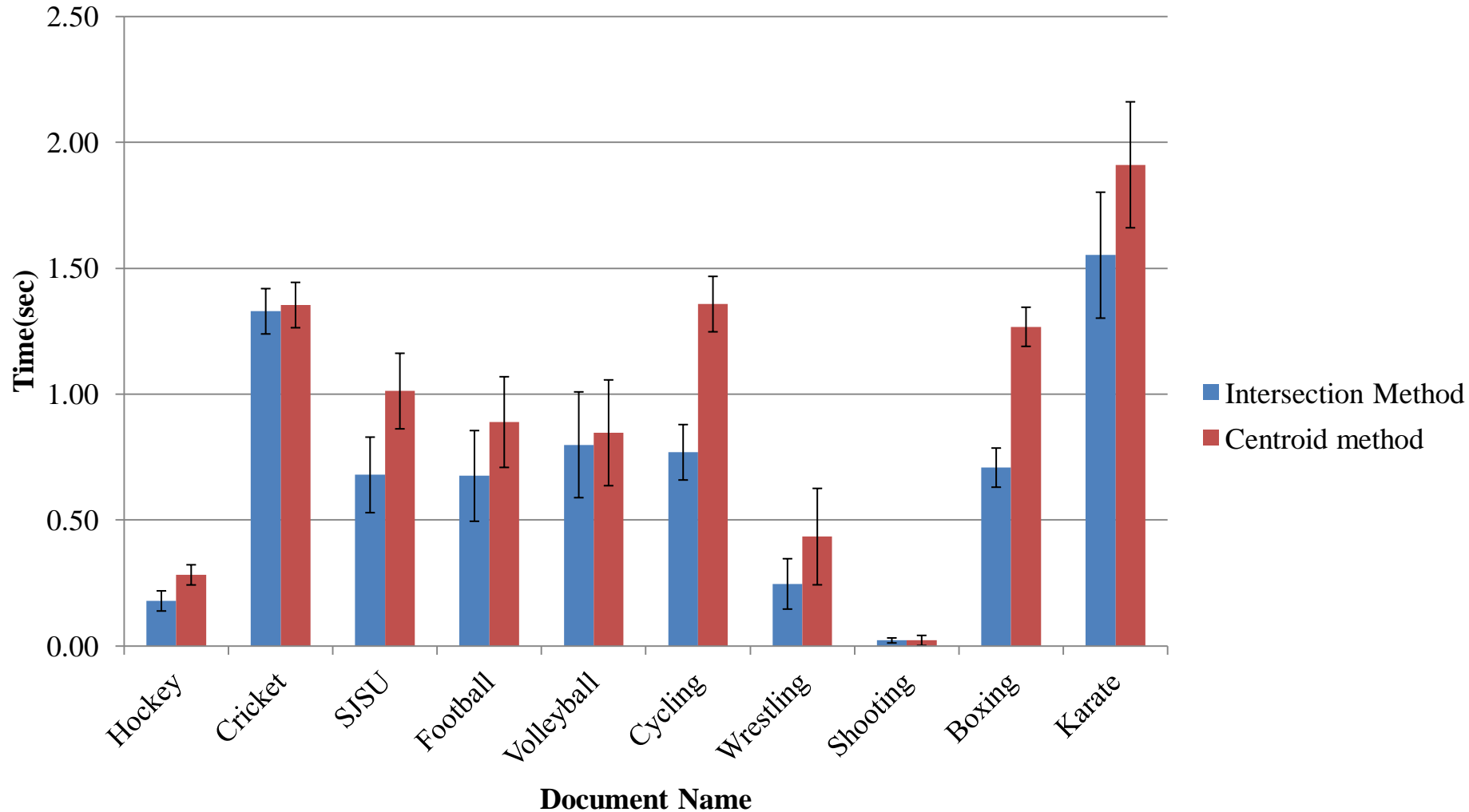
Similarity scores between Human generated summary and summarizer generated summary



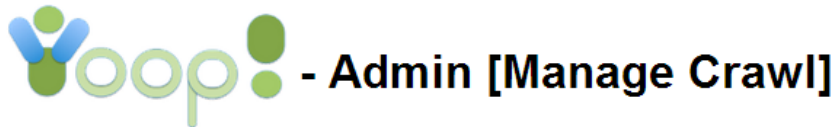
Time required to generate the summary for each of the three methods



Time required to generate the summary for Intersection and centroid method



Summarization Feature In Yioop



- Account Access**
- Manage Account
- Manage Users
- Manage Roles
- Manage Groups

- Blogs, Mixes**
- My Group Feeds
- Mix Crawls

- Crawls**
- Manage Crawl
- Manage Classifiers
- Page Options
- Results Editor
- Search Sources

- System Settings**

Edit Crawl Options [Back](#)

[Web Crawl](#) [Archive Crawl](#)

Get Crawl Options From: Use options below ▾

Crawl Order: Page Importance ▾

Restrict Sites By Url:

Summarizer: Centroid ▾
Basic
Allowed To Crawl: Centroid

Word Cloud


- Word cloud can be defined as
“a visual representation of keywords from the webpage”^[4]
- These keywords describes the complete webpage.
- Helps us to get the overall picture of the complete webpage.
- Also has the hyperlink associated with them to search that particular word on the Yioop search engine.

Word Cloud

PHP Search Engine - Yioop x

localhost/yioop/?YIOOP_TOKEN=6vmjn1TEkTE%7C1399015697&its=1398458393&q=earth

Settings



earth Search

0.31532 seconds. Showing 1 - 10 of 2751

[Earth](#)

en.wikipedia.org/wiki/Earth

Word cloud: sun earth moon sciences international

Earth ... is sometimes referred to as the world or the Blue Planet. **Earth** ... appeared on its surface within its first billion years. **Earth's** ... as the formation of the ozone layer, which together with **Earth's**
[Cached](#). [Similar](#). [Inlinks](#). [IP:198.35.26.96](#). Score:10.1

[Ecliptic](#)

en.wikipedia.org/wiki/Ecliptic

Word cloud: sun system celestial planets path

against the background stars as seen from the orbiting **Earth**. ... the path the Sun appears to trace through the stars. The **Earth** ... this path, which is coplanar with both the orbit of the **Earth**
[Cached](#). [Similar](#). [Inlinks](#). [IP:198.35.26.96](#). Score:9.83

[Earth_\(classical_element\)](#)

[en.wikipedia.org/wiki/Earth_\(classical_element\)](http://en.wikipedia.org/wiki/Earth_(classical_element))

Word cloud: black represented bile mother associated

Earth_ ... classical_element . **Earth** ... tradition. Image:La Terre Benoit Massou original.jpg**Earth** ... cults, and chthonic underworld deities, the element of **earth**
[Cached](#). [Similar](#). [Inlinks](#). [IP:198.35.26.96](#). Score:9.76

Multi-language support

PHP Search Engine - Yioop x

localhost/yioop/?YIOOP_TOKEN=ACie2kmS4CY%7C1398655102&its=1398192223&q=+ 维基百科

Settings



维基百科 Search

Search: [维基百科](#) 0.099683 seconds. Showing 1 - 10 of 401

[维基百科:关于 - 维基百科, 自由的百科全书](#)
zh.wikipedia.org/wiki/Wikipedia:关于 Word cloud: 編輯 维基百科 網路 百科全書 查看
维基百科:关于. ... 维基百科, 自由的百科全书.
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:10.1

[维基百科, 自由的百科全书](#)
zh.wikipedia.org/wiki/Wikipedia:首页 Word cloud: 拜登 公园 教师 世界地球日 美国
维基百科, 自由的百科全书. ... 维基百科:首页. 维基百科, 自由的百科全书. 跳转至: ... 维基百科:首页. 维基
百科, 自由的百科全书. 跳转至: 导航、搜索. 维基百科. 海納百川, 有容乃大.
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.44

[维基百科:特色条目 - 维基百科, 自由的百科全书](#)
zh.wikipedia.org/wiki/Wikipedia:特色條目 Word cloud: 公园 飓风伊莎贝尔 维基百科 植物 特色条目
维基百科:特色条目 ... - 维基百科, 自由的百科全书. 维基百科:特色条目. ... - 维基百科, 自由的百科全书. 维
基百科:特色条目. 维基百科, 自由的百科全书. (重定向自 Wikipedia:特色條目).
[Cached](#) [Similar](#) [Inlinks](#) [IP:198.35.26.96](#) Score:9.20

[维基百科:特色条目 - 维基百科, 自由的百科全书](#)
zh.wikipedia.org/wiki/Wikipedia:特色条目 Word cloud: 公园 飓风伊莎贝尔 维基百科 植物 特色条目
维基百科:特色条目 ... - 维基百科, 自由的百科全书. 维基百科:特色条目. ... - 维基百科, 自由的百科全书. 维
基百科:特色条目. 维基百科, 自由的百科全书. 跳转至: 导航、搜索. 中文维基百科特色條目.

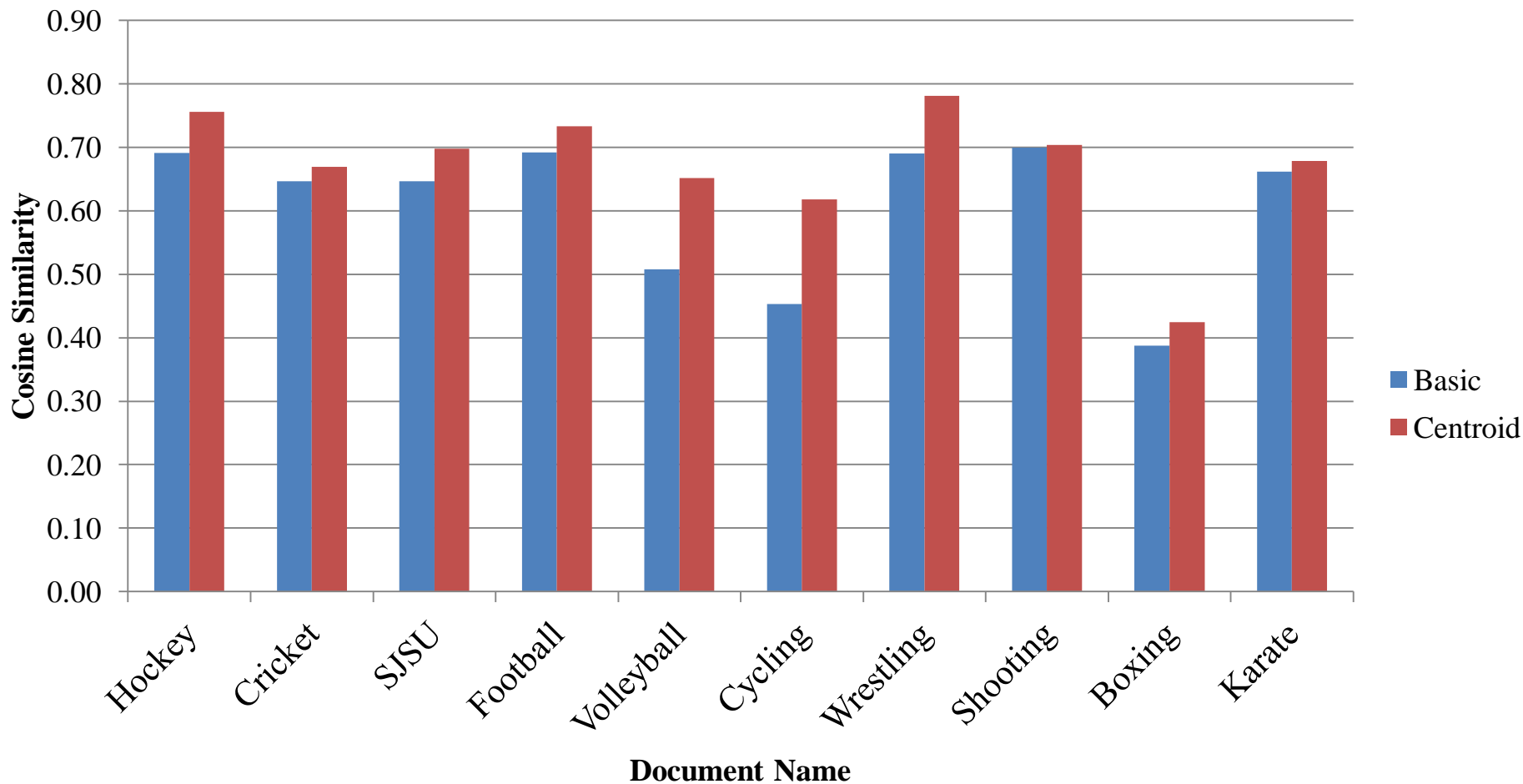
Experiments

- Evaluated the summarizer on basis of quality of the generated summary and time required to crawl 10,000 documents
- Comparison between the centroid summarizer and basic summarizer.

Quality of summary

- Main purpose of the project
- Ran basic and centroid summarizer on same set of documents
- Calculated the cosine similarity between the human generated summary and the summary generated by our two summarizers.

Cosine similarity of summary generated by Basic and Centroid summarizer with a human generated summary



Human generated summary

Football refers to a number of sports that involve, to varying degrees, kicking a ball with the foot to score a goal. The various codes of football share certain common elements. Players in American football, Canadian football, rugby union and rugby league take-up positions in a limited area of the field at the start of the game. The Ancient Greeks and Romans are known to have played many ball games, some of which involved the use of the feet. Games played in Mesoamerica with rubber balls by indigenous peoples are also well-documented as existing since before this time, but these had more similarities to basketball or volleyball, and since their influence on modern football games is minimal, most do not class them as football. A game known as "football" was played in Scotland as early as the 15th century: it was prohibited by the Football Act 1424 and although the law fell into disuse it was not repealed until 1906. King Henry IV of England also presented one of the earliest documented uses of the English word "football".

Summary generated by basic summarizer

Various forms of football can be identified in history, often as popular peasant games. Contemporary codes of football can be traced back to the codification of these games at English public schools in the eighteenth and nineteenth centuries. [2] [3] The influence and power of the British Empire allowed these rules of football to spread to areas of British influence outside of the directly controlled Empire, [4] though by the end of the nineteenth century, distinct regional codes were already developing: Gaelic Football, for example, deliberately incorporated the rules of local traditional football games in order to maintain their heritage. [5] In 1888, The Football League was founded in England, becoming the first of many professional football competitions. During the twentieth century, several of the various kinds of football grew to become among the most popular team sports in the world. [6] .. The various codes of football share certain common elements.

Summary generated by centroid summarizer

Football. Football refers to a number of sports that involve, to varying degrees, kicking a ball with the foot to score a goal. The most popular of these sports worldwide is association football, more commonly known as just "football" or "soccer". Unqualified, the word football applies to whichever form of football is the most popular in the regional context in which the word appears, including association football, as well as American football, Australian rules football, Canadian football, Gaelic football, rugby league, rugby union, and other related games. Association football, Australian rules football and Gaelic football tend to use kicking to move the ball around the pitch, with handling more limited. In most codes, there are rules restricting the movement of players offside, and players scoring a goal must put the ball either under or over a crossbar between the goalposts. It is widely assumed that the word "football" or "football" references the action of the foot kicking a ball.

Effect on Crawl time

- Crawled 10,000 pages by basic and centroid summarizer.
- Used Wikipedia database to make sure we are crawling same set of pages.
- Crawling 10,000 pages with basic summarizer took 28 minutes while crawling the same set of pages with centroid summarizer took 39 minutes.

HipHop Compiler for PHP

- HipHop is a static compiler developed by Facebook
- The standard implementation of PHP is an interpreter to support all the dynamic features of PHP.
- This interpreter is called Zend which is a bytecode interpreter.

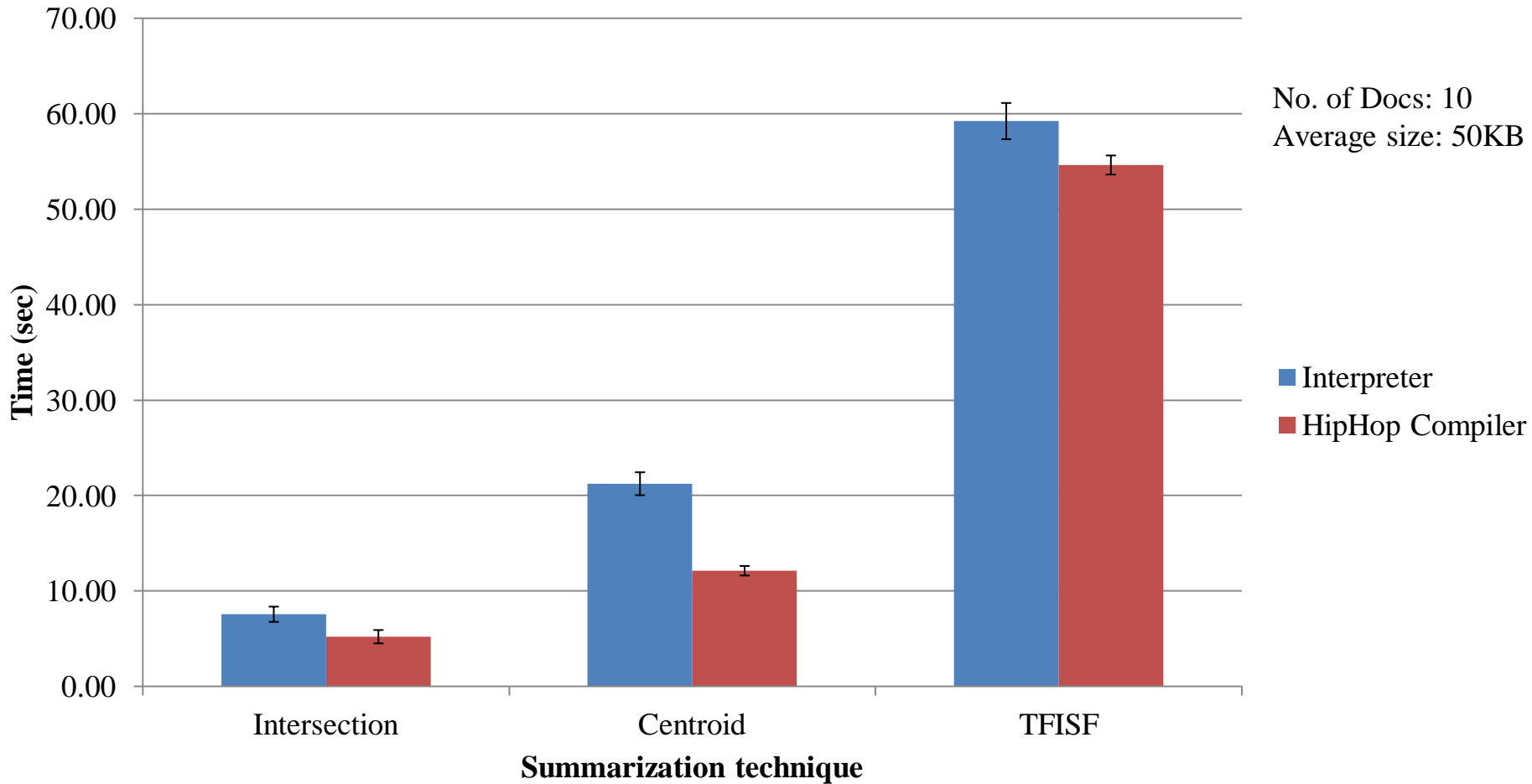
Differences between Zend Interpreter and HipHop compiler

- HipHop compiler needs all source code to be known in advance
- HipHop doesn't support all features of the PHP like dynamic code evaluation
- HipHop analyzes, compiles and loads all the symbols in advance
- A small amount of change in a code can result in rebuilding the system

Experiment

- Ran all three summarizer methods on HipHop compiler and compared the time required to generate the summary with time required on the Zend interpreter
- Document set : 10 documents
- Average document size : 50KB

Comparison between using Interpreter and Compiler for running summarizers



Conclusion

- Intersection method is the fastest method among the three.
- The centroid method generates the best summary among the three.
- Integrated the centroid summarizer into the Yioop.
- Stop words remover for other languages.

References

1. Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. *In Proceedings of EMNLP* (Vol. 4, No. 4).
2. Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6), 919-938.
3. Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade, Formulation of document summarization as a 0–1 nonlinear programming problem, *Computers & Industrial Engineering*, Volume 64, Issue 1, January 2013, Pages 94-102, ISSN 0360-8352
4. Halvey, M. J., & Keane, M. T. (2007, May). An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1313-1314). ACM.
5. Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). *Information retrieval: Implementing and evaluating search engines*. Mit Press.
6. Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5, 361-397.

Thank you