SIAM-IMA Etymo workshop - Keyword Extraction

Steven Elsworth

June 13, 2018

Keyword Extraction

Example text

List of keywords

Example Data

Title: Compatibility of systems of linear constraints over the set of natural numbers

Abstract: Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

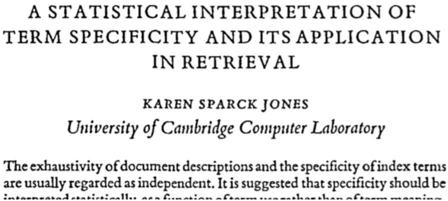
Manually assigned keywords: linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

Neural nets

We do not have the training data.

- `https://www.worldcat.org`
- `https://www.sciencedirect.com/science/article/pii/S0957417417308473`
- `https://www.worldscientific.com/doi/abs/10.1142/S0218213004001466`
- `https://epubs.siam.org/doi/abs/10.1137/17M1119901`

TFIDF: Term Frequency, Inverse Document Frequency



A STATISTICAL INTERPRETATION OF
TERM SPECIFICITY AND ITS APPLICATION
IN RETRIEVAL

KAREN SPARCK JONES
University of Cambridge Computer Laboratory

The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning

Published in 1972, led to TF-IDF .

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

$$IDF(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term t}}\right).$$

$$Value = TF * IDF$$

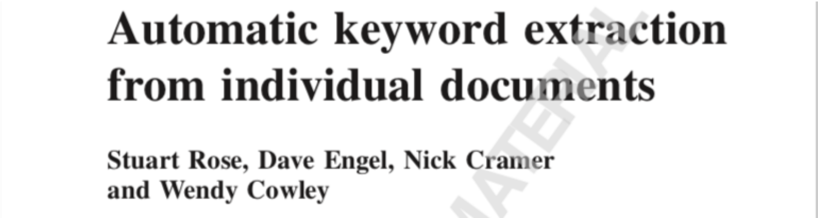
TFIDF Example

`http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

- lowercase
- stopwords
- n-gram range
- idf (smoothing)
- document frequency range

'systems' : $\frac{4}{74\log(1)}$, , 'solutions' : $\frac{4}{74\log(1)}$, 'minimal' : $\frac{4}{74\log(1)}$, 'types' : $\frac{4}{74\log(1)}$, 'considered' : $\frac{3}{74\log(1)}$, 'set', : $\frac{3}{74\log(1)}$, 'set solutions' : $\frac{3}{74\log(1)}$, ...

RAKE: Rapid Automatic Keyword Extraction



Automatic keyword extraction from individual documents

Stuart Rose, Dave Engel, Nick Cramer and Wendy Cowley

'An unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents'

'The input parameters for RAKE comprise a list of stop words, a set of phrase delimiters, and a set of word delimiters.'

RAKE Example

- **Document:** 'A matrix is a rectangular array of numbers, symbols, or expressions, arranged in rows and columns.'
- **Stop words:** Fox, Christopher. "A stop list for general text." ACM SIGIR Forum. Vol. 24. No. 1-2. ACM, 1989. Example

  = ' a, about, above, across, after, again, against, all, almost, alone, along, already, also, although, always, among, an, ...'
- **Phrase delimiters:** '.', ',', '!', ':', ';'
- **Word delimiters:** ' ', ' ', ' ', ' ', ' ', ...

RAKE: Extracted Keywords

https://github.com/zelandiya/RAKE-tutorial

compatibility, systems, linear constraints, set, natural numbers, criteria, compatibility, system, linear diophantine equations, strict inequations, nonstrict inequations, upper bounds, components, minimal set, solutions, algorithms, minimal generating sets, solutions, systems, criteria, corresponding algorithms, constructing, minimal supporting set, solving, systems, systems

RAKE: Extracted Keywords

`https://github.com/zelandiya/RAKE-tutorial`

compatibility, systems, linear constraints, set, natural numbers, criteria, compatibility, system, linear diophantine equations, strict inequations, nonstrict inequations, upper bounds, components, minimal set, solutions, algorithms, minimal generating sets, solutions, systems, criteria, corresponding algorithms, constructing, minimal supporting set, solving, systems, systems

$$\text{score} = \frac{\texttt{deg}(w)}{\texttt{freq}(w)}$$

RAKE: Extracted Keywords

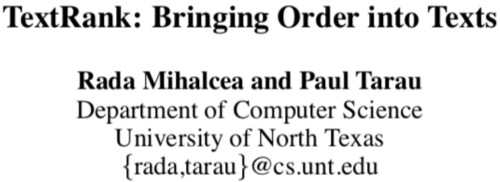
`https://github.com/zelandiya/RAKE-tutorial`

compatibility, systems, linear constraints, set, natural numbers, criteria, compatibility, system, linear diophantine equations, strict inequations, nonstrict inequations, upper bounds, components, minimal set, solutions, algorithms, minimal generating sets, solutions, systems, criteria, corresponding algorithms, constructing, minimal supporting set, solving, systems, systems

$$\text{score} = \frac{\mathtt{deg}(w)}{\mathtt{freq}(w)}$$

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

Graph Based Approach



TextRank: Bringing Order into Texts

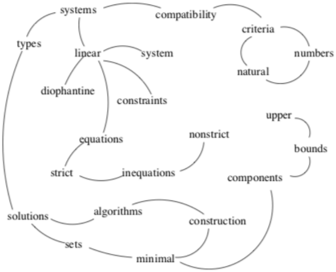
Rada Mihalcea and Paul Tarau
Department of Computer Science
University of North Texas
{rada,tarau}@cs.unt.edu

1. Identify text units that best define the task at hand, and add them as vertices in the graph.
2. Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.
3. Graph-based ranking algorithm.
4. Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions.

Graph Based Approach Example

`https://github.com/davidadamojr/TextRank`

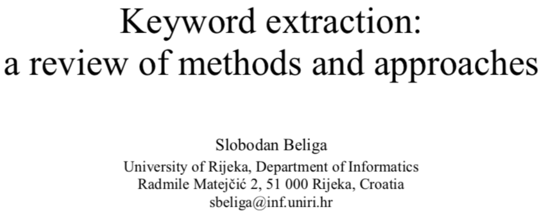
Construct co-occurrence relation, two vertices are connected if the words appear within a window of maximum words, where can be set anywhere from 2 to 10 words.



Apply graph ranking algorithm:
numbers (1.46), inequations (1.45), linear (1.29), diophantine (1.28), upper (0.99), bounds (0.99), strict (0.77), ...

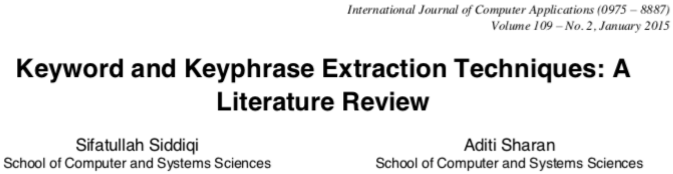
Postprocessing.

Literature Review



Keyword extraction: a review of methods and approaches

Slobodan Beliga
University of Rijeka, Department of Informatics
Radmile Matejčić 2, 51 000 Rijeka, Croatia
sbeliga@inf.uniri.hr



International Journal of Computer Applications (0975 – 8887)
Volume 109 – No. 2, January 2015

Keyword and Keyphrase Extraction Techniques: A Literature Review

Sifatullah Siddiqi
School of Computer and Systems Sciences

Aditi Sharan
School of Computer and Systems Sciences