



Enhancing Text Clustering by Leveraging Wikipedia Semantics

BAŞAK ÜNSAL

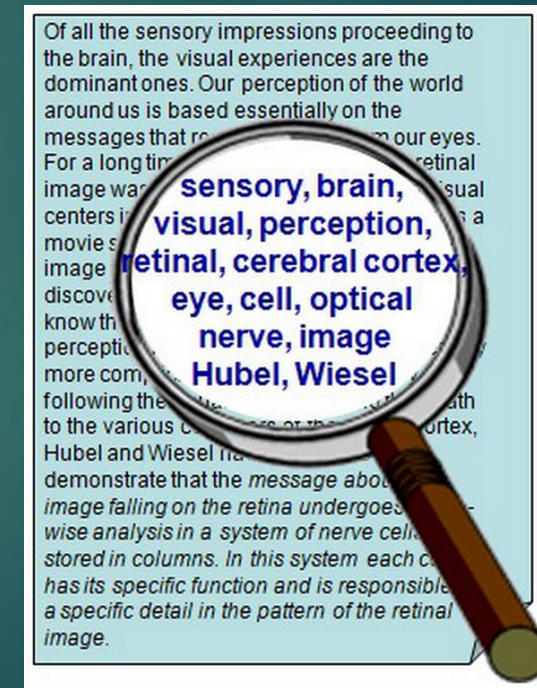
ÖZGE YALÇINKAYA

Content

- ▶ Motivation
- ▶ Related Works
- ▶ Methodology
- ▶ Outcomes
- ▶ Opinions

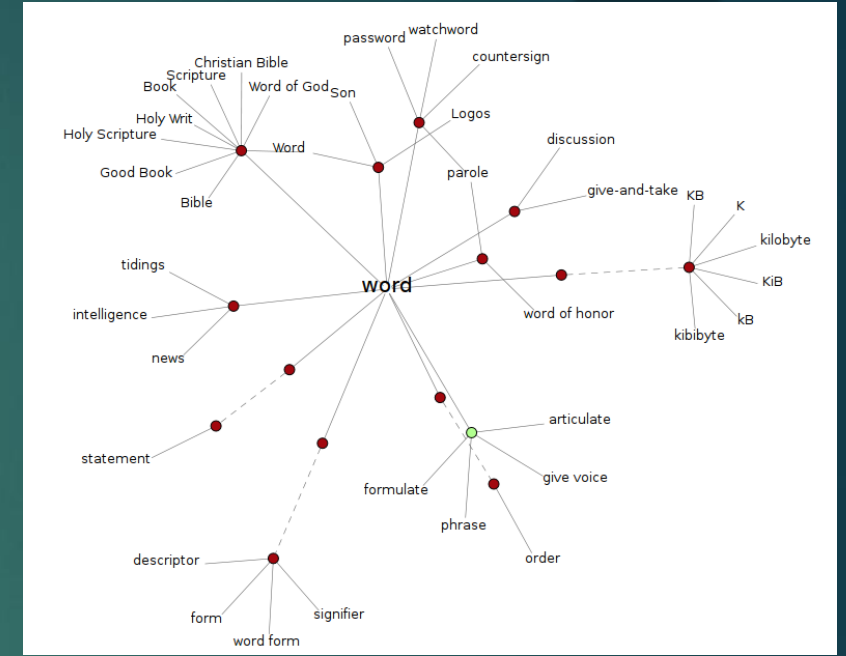
Motivation

- ▶ Text clustering algorithms have importance to provide better document organization.
- ▶ Traditional one is «bag of words» (BOW).
 - ▶ Looks frequencies of terms in a document.
 - ▶ Ignores the semantic relationships between key terms.
 - ▶ While clustering, BOW does not reflect the distance of two documents accurately.
- ▶ Finding an accurate distance measure is crucial to improve text clustering.



Related Works

- ▶ Buenaga Rodriguez et al. [1] and Hotho et al. [2] integrated the WordNet resource to categorize documents
 - ▶ WordNet groups English words into sets of synonyms called synsets
 - ▶ However, it has limited coverage and disambiguation is not enough
- ▶ Gabrilovich et al. [3] apply feature generation using Wikipedia
 - ▶ Background knowledge based features helps
 - ▶ Wikipedia has less noise data
 - ▶ They don't use the relations in Wikipedia such as hyponym and synonym.



[1] M. de Buenaga Rodriguez, J. M. G. Hidalgo, and B. DiazAgudo. Using WordNet to complement training information in text categorization.

[2] A. Hotho, S. Staab and G. Stumme. Wordnet improves text document clustering.

[3] E. Gabrilovich and S. Markovitch. Feature Generation for Text Categorization Using World Knowledge.

Methodology

1. Wikipedia Thesaurus

- ▶ A concept thesaurus is created based on Wikipedia's semantic relations such as,
 - ▶ Synonym
 - ▶ There is only one article for each concept. (US,USA)
 - ▶ Polysemy
 - ▶ Lists all possible meanings for a term. (Puma)
 - ▶ Hypernym
 - ▶ Each article can belong more than one concept.
 - ▶ Associative relations
 - ▶ Relatedness of hyperlinks within articles.

US (disambiguation)

From Wikipedia, the free encyclopedia
(Redirected from [Us](#))

US or **U.S.** usually refers to the [United States of America](#), a country in North America.

US, **U.S.**, **Us**, **us**, or **u.s.** may also refer to:

Language [\[edit\]](#)

- The objective case of [we](#) in English

Geography [\[edit\]](#)

- [Us](#), Val-d'Oise, France

Puma

From Wikipedia, the free encyclopedia

Puma may refer to:

Animals [\[edit\]](#)

- [Puma \(genus\)](#), the genus containing the cougar and the [jaguarundi](#)
- [Cougar](#), a large cat also known as a puma, mountain lion, panther or catamount

Companies [\[edit\]](#)

- [Puma SE](#), a German shoe and sportswear company
- [Puma Energy International](#), a petrol distribution company, subsidiary of [Trafigura Beheer B.V.](#)

Vehicles [\[edit\]](#)

Sports cars

- [Puma \(car\)](#), a Brazilian brand of sports cars
- [Puma \(kit-car\)](#), an Italian brand of dune buggy and sports kit-car
- [Personal Urban Mobility and Accessibility](#), a prototype electric vehicle designed by Segway and General Motors
- [Ford Puma](#), a sports car

Combat vehicles

- [Puma \(AFV\)](#), an Italian family of armoured fighting vehicles
- [Puma \(IFV\)](#), a German infantry fighting vehicle
- [Puma \(IDF\)](#), an Israeli combat engineering vehicle
- [PLUMA M26-15](#), a South African armoured personnel carrier

Methodology

2. Improving Text Clustering using Wikipedia Thesaurus

- ▶ Traditional Text Similarity Measure:
 - ▶ Extract BOW representations,
 - ▶ Look similarity. Articles are similar, if they share common terms.
 - ▶ Calculate S_{TFIDF} .
- ▶ Traditional Text Representation Enrichment Strategies:
 - ▶ Previous ones enriched with text representation with external resources
 - ▶ Wordnet and Open Directory Project
 - ▶ Generates new features (synonym or hypernym)
 - ▶ Appends to original document and constructs new vector.

Proposed Method

- ▶ Map terms in documents to Wikipedia concepts.
 - ▶ Building phrase index
- ▶ Enriching Similarity Measure with Hierarchical Relation:
 - ▶ Each concept can belong one or more categories:
 - ▶ $cate_c = \{cate_{c1}, cate_{c2}, \dots, cate_{cm}\}$
 - ▶ Similarity is:
 - ▶ $S = (1 - \alpha)S_{TFIDF} + \alpha S_{cate}$
- ▶ Synonym and Associative Relation:
 - ▶ $asso_c = \{(cr_1, w_1), (cr_2, w_2), \dots\}$ where cr_1 is the related concept and w_1 is relatedness.
 - ▶ Similarity is:
 - ▶ $S = (1 - \beta)S_{TFIDF} + \beta S_{asso}$
- ▶ Combination of them:
 - ▶ $S_{comb} = (1 - \alpha - \beta)S_{TFIDF} + \alpha S_{cate} + \beta S_{asso}$

Outcomes

- ▶ Purity measure is applied.
- ▶ Three baselines are used.
 - ▶ K-Means with traditional text document similarity measure (Base 1)
 - ▶ K-Means with document representation (Gabrilovich's version) (Base 2)
 - ▶ K-Means with text document representation (Hotho's version) (Base 3)

	Reuters			OHSUMED		
	Purity	Inverse	Impr	Purity	Inverse	Impr
<i>BASE1</i>	0.603	0.544		0.414	0.343	
<i>BASE2</i>	0.605	0.548	0.53%	0.427	0.354	3.17%
<i>BASE3</i>	0.607	0.556	1.43%	0.435	0.358	4.72%
<i>HR</i>	0.604	0.547	0.36%	0.459	0.388	12.0%
<i>SAR</i>	0.652	0.593	8.57%	0.438	0.359	5.23%
<i>COB</i>	0.655	0.598	9.28%	0.449	0.381	9.77%

Opinions

- ▶ Their work is useful for clustering documents since,
 - ▶ It does not look only the frequency but also the semantics
 - ▶ They use Wikipedia which has a good coverage and semantic relations
- ▶ They have good results compared to other methods.