Automatic Ontology Population from News

ATCM Lab Work

Hernani Costa

hpcosta@dei.uc.pt

Supervisor: Alexandre Miguel Pinto

ampinto@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

Coimbra, 13 July, 2012

Introduction

We are experiencing a huge growth in the available information, most of the time provided as natural language text (e.g., daily news)

Keyphrases provide a brief summary of a document's contents

- usually chosen manually
- less prone to errors

- hardly repeatable
- time-consuming
- sometimes subjective

Introduction

Methodology & Goals

A system capable of:

- extract keyphrases automatically from heterogeneous Web sources
- representing them in an ontology
- identify main topics in the ontology

The goals are:

- ascertain how well-suited are machine learning approaches for keyphrases extraction task
- represent the resulting knowledge into a well-defined knowledge base, that will provide support to a News Recommender System

Keyphrase Extraction Approaches

Four existing approaches:

- Statistical
  - simple (e.g., TF-IDF (Salton and Buckley (1988)))
  - non-linguistic features
- Linguistic
  - add value, e.g., by filtering 'undesirable' words (van der Plas et al. (2004))
  - normally uses mixed methods (linguistic + statistics)
- Machine Learning
  - usually uses supervised learning (Witten et al. (1999))
- Hybrid
  - combines the aforementioned methods (Medelyan (2009))

KEA & Neo4j

KEA[1] (**K**eyword **E**xtraction **A**lgorithm)

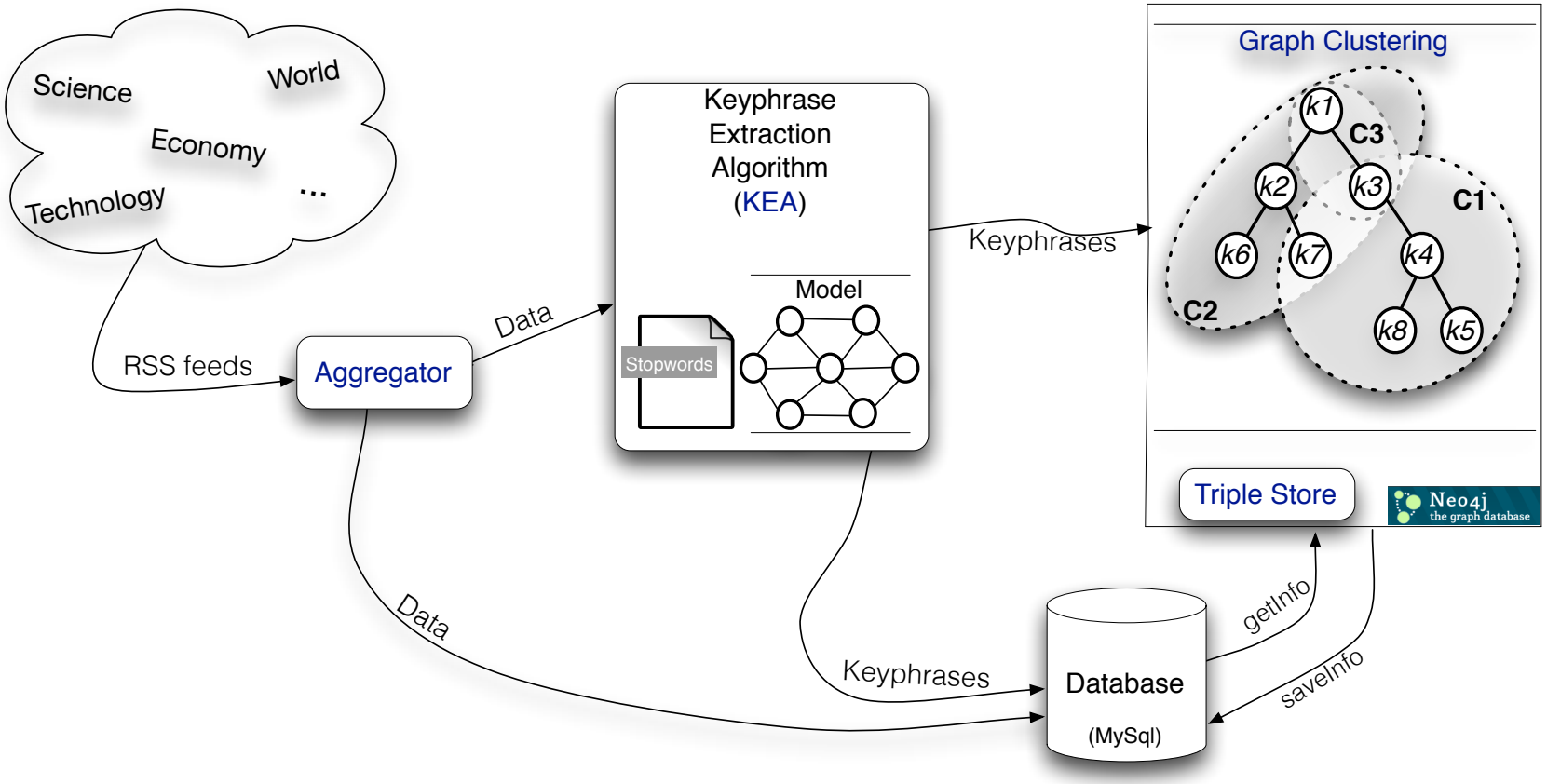
- includes a cut-down version of the Weka
- it can be either used
  - for free indexing
  - for indexing with controlled vocabulary

Neo4j[2]

- is a graph database
- nodes are interconnected with an undetermined number of relations between them

[1] http://www.nzdl.org/Kea
[2] http://neo4j.org

System Overview



KEA's Model Evaluation

Table: KEA's Evaluation.

	S	E	D(%)	T(%)	P(%)	R(%)	F_1(%)
Science	85	20	70.0	30.0	85.4	57.1	66.5
World	632	41	70.7	29.3	66.6	63.0	61.3
Economy	35	24	70.8	29.2	60.7	48.7	52.4
Technology	30	22	72.3	27.3	47.9	72.2	53.5
Macintosh	13	10	70.0	30.0	46.8	68.3	54.4
Video-DVD	26	16	68.8	31.2	37.5	56.4	43.3
Cinema	19	16	68.8	31.2	27.5	39.1	28.9
	Total		**Average**				
	840	**149**[3]	**70.3**	**29.7**	**53.2**	**57.8**	**51.3**

S - news in the system

E - news manually evaluated

D - % used in the training process

T - % used to test the algorithm

P - precision

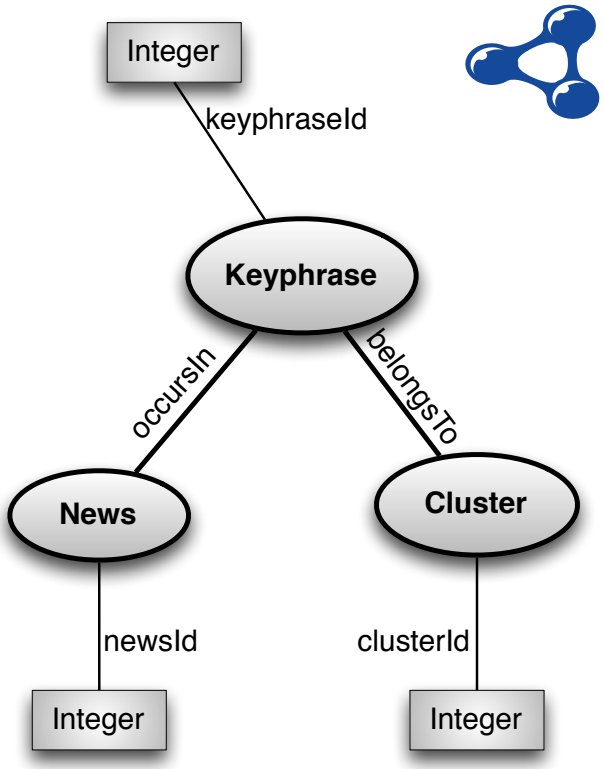
R - recall

F_1 - F measure

[3] Sample set has a confidence level of 95% and a confidence interval of 7.3%.

Ontology Definition & Population

Neo4j as a Triple Store



Ontology definition

- RDF shema created in Protégé[a]

Ontology Population

- triples
  - ★ t_n={$keyphrase_n$, occursIn, $news_n$}, where $_n \in N$)
- gathering news from one week (i.e., 7 days * 24 hours)
  - ★ 4683 nodes (keyphrases)
  - ★ 18569 edges (connections between the keyphrases and the news)

[a] http://protege.stanford.edu

Graph-Clustering

Loading ...

Concluding Remarks

- Some scores in the extraction process revealed F_1 scores higher than 60%
- Graph-Clustering can be considered as a successful proof of concept
- The prototype can be seen as a first step in the construction of an automatic keyphrase extract system
- Made for Portuguese, but easily adapted to other languages

Prototype system capable of:

- extract keyphrases from different sources
- formally represent the extracted knowledge into an ontology
- identify main topics

Future Work

- Extraction algorithm can be improved by using
  - pre-filtering
    - increasing the stopword list
    - the usage of a POS tagger + a grammar to create rules before the extraction process (Costa (2011))
  - post-filtering
    - identify the keyphrases POS (e.g., to discard verbs)
    - rate the keyphrases to improve the algorithm accuracy (Costa et al. (2010))
    - taking advantage of other resources, e.g., Onto.PT[4] or BDpedia[5]
- Use more news to study the effect in the system accuracy
- Explore other keyphrase extraction algorithms and perform a comparison between them

[4] http://ontopt.dei.uc.pt
[5] http://dbpedia.org

References I

Costa, H. (2011). *Automatic Extraction and Validation of Lexical Ontologies from text: Creating Lexical Ontologies from text*. LAP LAMBERT Academic Publishing.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2010). The Impact of Distributional Metrics in the Quality of Relational Triples. In *Proc. ECAI 2010, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.

Medelyan, O. (2009). *Human-competitive automatic topic indexing*. PhD thesis, University of Waikato, New Zealand, Waikato, New Zealand.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an Int. Journal*, 24(5):513–523.

van der Plas, L., Pallotta, V., Rajman, M., and Ghorbel, H. (2004). Automatic Keyword Extraction from Spoken Text. A Comparison of two Lexical Resources: the EDR and WordNet. *CoRR*, cs.CL/0410062.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). KEA: practical automatic keyphrase extraction. In *Proc. 4th ACM Conf. on Digital Libraries*, DL'99, pages 254–255, NY, USA. ACM.


Recommender Systems
Recall
Clustering Algorithms
F1
Precision
Neo4j
Evaluation
Aggregation
Yahoo Content Analysis
Daily News
Weka
Linguistic Approaches
Hybrid Approaches
KEA
Statistic Approaches
Automatic Extraction
Artificial Intelligence
Machine Learning Approaches
Graph Clustering
AlchemyAPI
Tools
Semantic Web
Natural Language Processing
Gephi
Clustering
System Prototype
Ontologies
News
Automatic Population
Portuguese Language
Libraries
Information Extraction