Semantic-Aware Expert Partitioning

Veselka Boeva, Liliana Boneva, Elena Tsiporkova

Technical University of Sofia
Department of Computer Systems & Technologies
Plovdiv, Bulgaria

AIMSA 2014
Varna, Bulgaria

Expertise Retrieval

- Finding the right person in an organization with the appropriate skills and knowledge is often crucial to the success of projects being undertaken.
- **Expert finders** are usually integrated into organizational information systems, such as knowledge management systems, recommender systems, and computer supported collaborative work systems.

- Initial approaches propose tools that rely on people to self-assess their skills against a predefined set of keywords.
- Later approaches try to find expertise in specific types of documents, such as e-mails or source code.
- Systems that index and mine published intranet documents as sources of expertise evidence are also proposed.
- In the recent years, research on identifying experts from online data sources has been gaining interest.

Possible applications

- Identification of experts in a particular technological domain, *e.g.* for the purpose of *technology scouting*.
- *Partner matching* for research proposals.
- Visualization of research activities and experts within geographical regions, *e.g.* in the context of *technology brokerage*.
- *. . .*

Expertise Retrieval Tasks

- **Expert finding** is the task of finding experts given a topic describing the required expertise.
- **Expert profiling** is the task of returning a list of topics that a person is knowledgeable about.

Clustering Analysis

- **Clustering analysis** is a process that partitions a set of objects into clusters in such a way that objects from the same cluster are similar and objects from different clusters are dissimilar.

- **Document clustering** is a widely studied problem with many applications such as document organization, browsing, summarization, classification.

Clustering of Experts

- The cluster hypothesis for **document retrieval** states that similar documents tend to be relevant to the same request.
- In the context of **expertise retrieval** the clustering hypothesis can be re-stated that similar people tend to be experts on the same topics.

Semantic-Aware Expert Partitioning

- Construction of Expert Profiles
  - Each expert is represented by lists of keywords, extracted from the available information about his/her expertise.
- Clustering of Topics (Keywords)
  - A common set of all different keywords is formed by pooling the keywords of all the expert profiles.
  - The semantic distance between each pair of keywords is calculated and the keywords are partitioned.
- Clustering of Experts
  - Each expert is represented by a vector of membership degrees of the expert to the different clusters of keywords.
  - The Euclidean distance between each pair of vectors is calculated and the experts are clustered.

Construction of Expert Profiles:

- The data needed for constructing the expert profiles can be extracted from various Web sources, *e.g., LinkedIn, the DBLP library, Microsoft Academic Search, Google Scholar Citation* etc.

- The **Stanford part-of-speech tagger** can be used to annotate the different words in the text collected for each expert with their specific part of speech.

- The annotated text can be reduced to a set of keywords (tags) by removing all the words tagged as articles, prepositions, verbs, and adverbs.

Construction of Expert Profiles:

- Only the nouns and the adjectives are retained and the final keyword set can be formed according to the following chunking algorithm:

  - *adjective-noun(s) keywords*: a sequence of an adjective followed by a noun is considered as one compound keyword *e.g.* ”supervised learning”;

  - *multiple nouns keywords*: a sequence of adjacent nouns is considered as one compound keyword *e.g.* ”mixture model”;

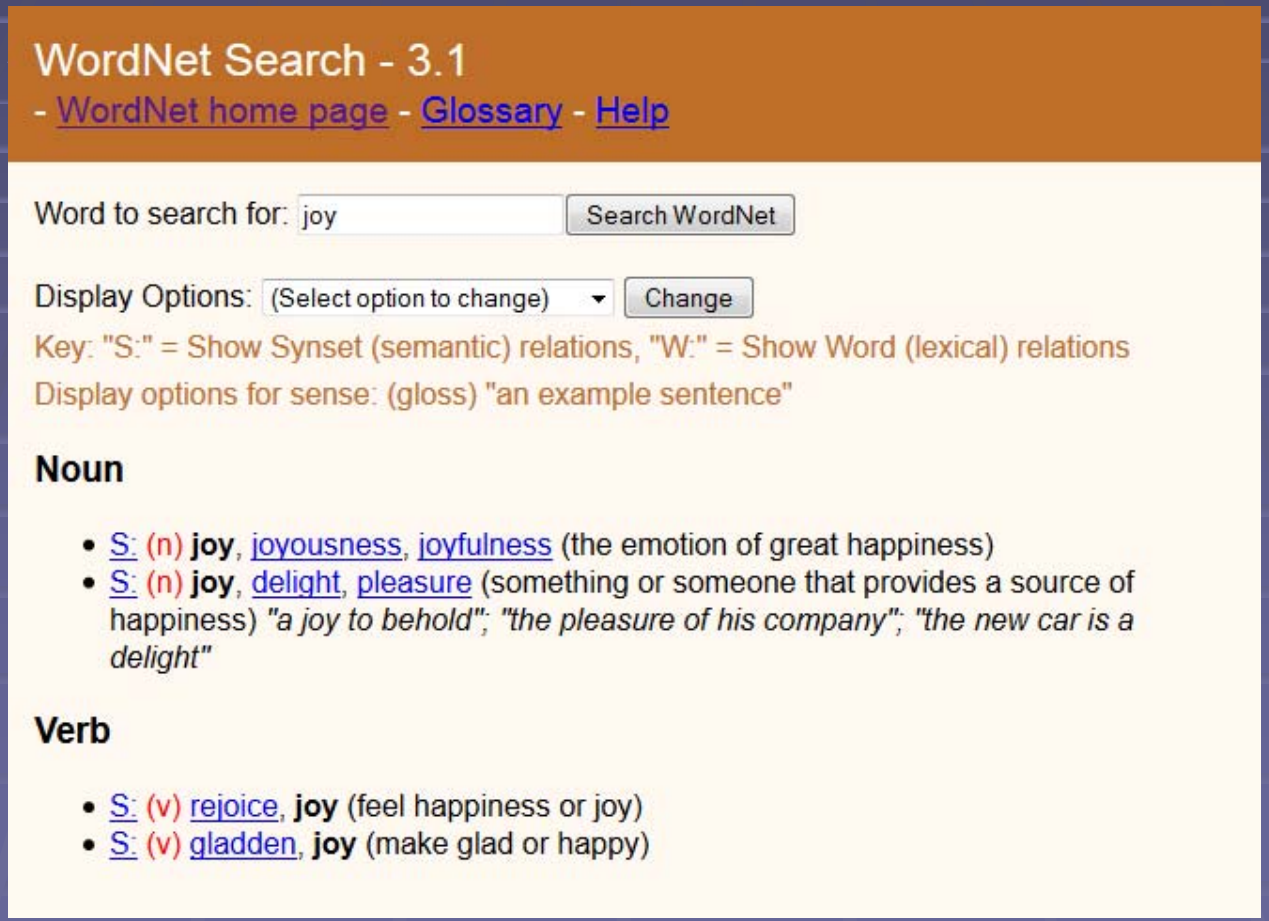
  - *single noun keywords*: each of the remaining nouns forms a keyword on its own.

Clustering of Keywords:

- A set of different keywords is formed by gathering all the keywords of all expert profiles.
- The semantic distance between each pair of keywords can be calculated by using the **WordNet** (a large lexical database of English).
  - Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.
  - Synsets are interlinked by means of conceptual-semantic and lexical relations.

Clustering of Keywords:

- Example of WordNet synsets:



Clustering of Keywords:

- The WordNet ontology constrains:
  - Initially, the WordNet networks for the four different parts of speech were not linked to one another and the noun network was the first to be richly developed.
  - Not all keywords representing the expert profiles are nouns.
  - The algorithms that can measure similarity between adjectives do not yield results for nouns.

Clustering of Keywords:

- Normalized similarity measure:

If m_i is an arbitrary similarity measure its normalized measure MN_i for any two keywords v and w can be calculated as follows:

$$MN_i(v, w) = m_i(v, w) / m_i(v, v),$$

where $m_i(v, v)$ gives the maximum possible score of m_i.

If m_i takes non-negative values, then MN_i takes values in [0, 1].

Clustering of Keywords:

- Combined similarity measure:

Our own normalized measure *MN* combined from *r* different similarity measures $m_1, m_2, \ldots, m_r$ can be computed as follows:

$$MN(v, w) = \alpha_1 MN_1(v, w) + \alpha_2 MN_2(v, w) + \ldots + \alpha_r MN_r(v, w),$$

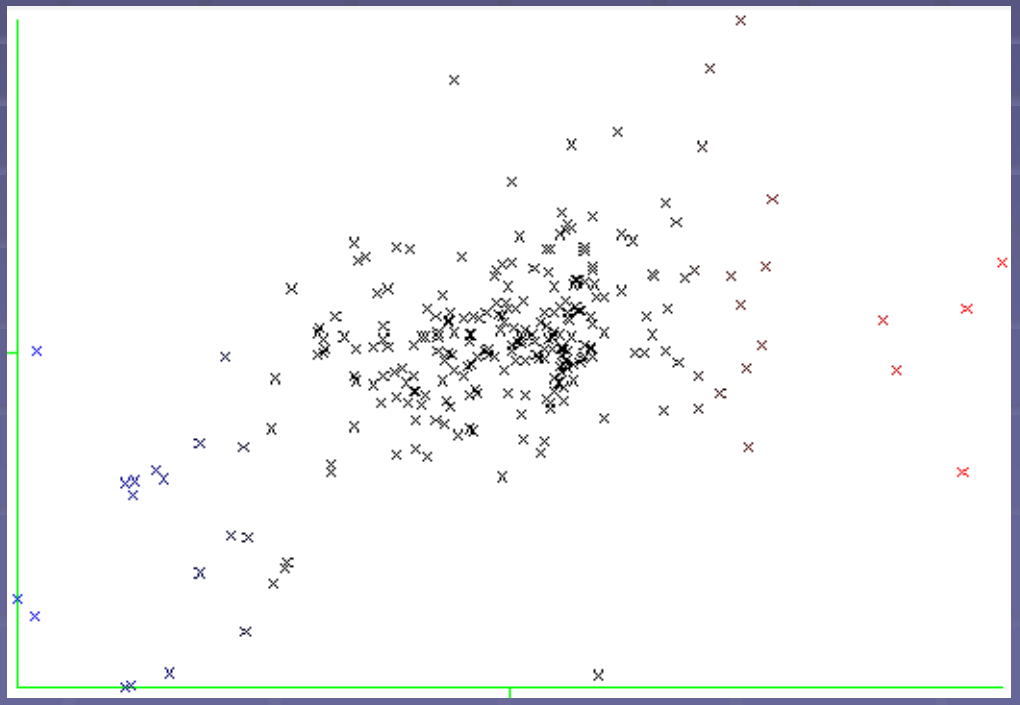
where α_i *denotes the weight of i-th measure and*

$$\alpha_1 + \alpha_2 + \ldots + \alpha_r = 1.$$

It is clear that *MN* takes values in [0, 1].

Clustering of Keywords:

- The keywords can be clustered by applying the ***k*-means** (or other partitioning) algorithm:
  - Decompose the data set into *k* disjoint clusters minimizing the within-cluster sum of distances
  - The cluster center is the mean data vector averaged over all objects in the cluster.



Clustering of Experts:

- Each expert is represented by a vector of membership degrees of the expert to the clusters of keywords.

The keywords are grouped into k clusters: $C_1, C_2, \ldots, C_k$.

b_{ij} is the number of keywords from the expert profile of expert i *that belong to cluster* C_j.

p_i is the total number of keywords in the expert profile of expert i.

Then each expert i can be represented by a vector

$$e_i = (e_{i1}, e_{i2}, \ldots, e_{ik}), \text{ where } e_{ij} = b_{ij} / p_i \ (j = 1, 2, \ldots, k).$$

- The Euclidean distance between each pair of vectors is calculated and the experts are grouped by applying the k-means or other clustering algorithm.

Initial Evaluation

- The test collection from a scientific conference (ITBAM 2011) devoted to information technology in bio- and medical informatics is used.
- For each topic, participants (53 in total) of the corresponding conference session are regarded as experts on that topic.
- A total of 5 topics (sessions) are created by the conference science committee.
- The names of researchers that are listed in the conference program on the sessions (topics) information are extracted. These researchers are considered as relevant experts and used as the ground truth in the validation.

Test Data

- The data needed for constructing the researcher expertise profiles are extracted from **Microsoft Academic Search**.
- A researcher profile is defined by a list of keywords used in the profile page of the author in question.
- Some of the keywords are multiple-word terms, *e.g.* ”Molecular Biology”, ”Data Mining”, ”Software Engineering”, ”Information Retrieval” etc.
- Not all the multiple-word terms are present in WordNet ontology. Therefore, these keywords have been divided into their constituting words.

Cluster Validation Measures

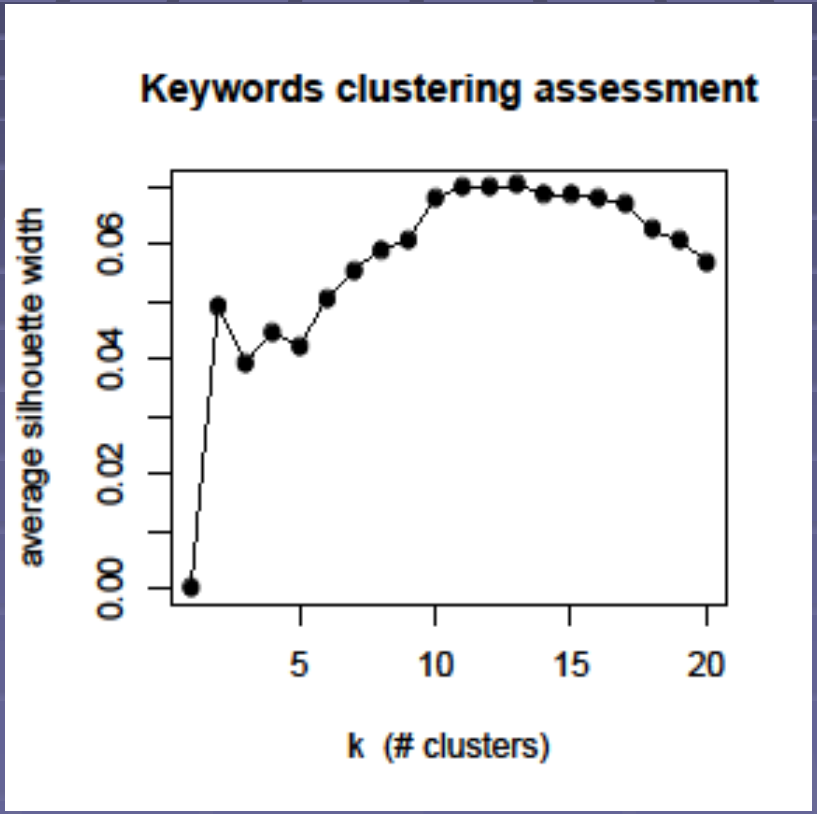
- **Cluster validation techniques** are designed to find the partitioning that best fits the underlying data.
- **Silhouette Index (SI)** is applied for assessing compactness and separation of a clustering solution.
  - the value of SI varies from -1 to 1 and should be **maximized**
- **SI** is also used as a validity index to identify the clustering scheme which best fits the test data.
- **F-measure** is used for evaluating the accuracy of the generated clustering solutions.
  - the maximum value of the F-measure is 1.

Implementation and Availability

- A free distributed **WordNet Similarity for Java** (WS4J) library has been used to measure the word similarity.
- The semantic relatedness algorithms implemented by the WS4J library have been used in our experiments.
- A normalization on all scores in order to obtain a final score in one and the same range has been performed.
- The weights are evenly distributed among the algorithms that produce a score for a given word pair.
- R scripts have been used to implement all the other experiments and to generate the result plots.

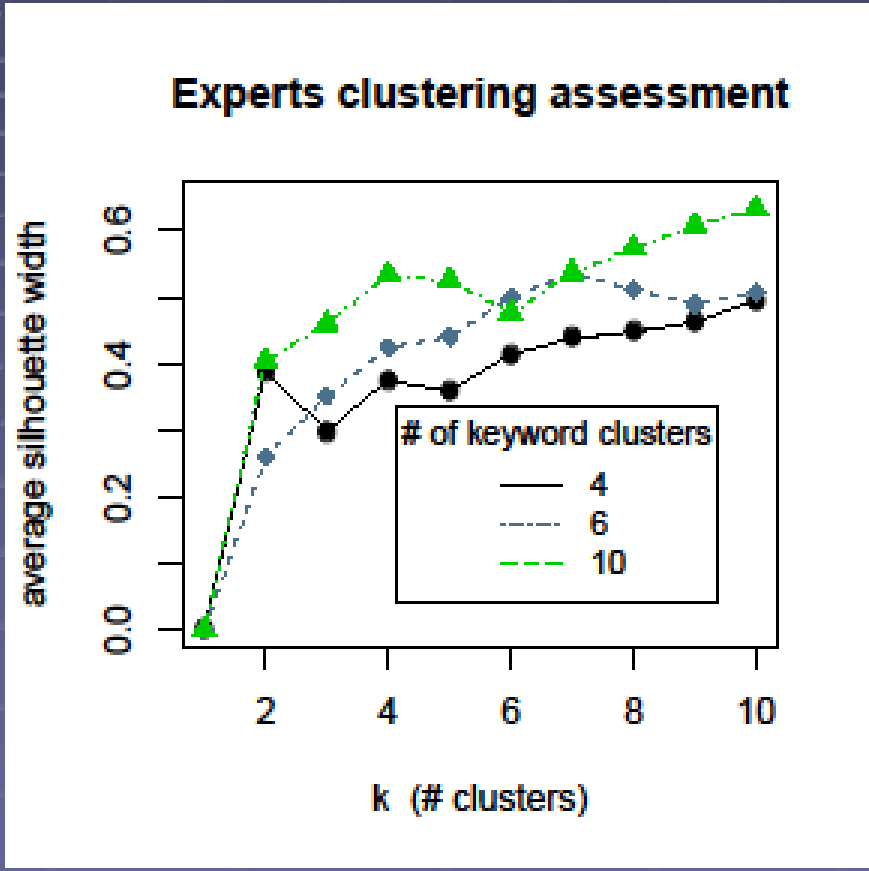
Experimental Results

A set of 44 different keywords is formed by gathering all the keywords of all 53 expert profiles.



SI values generated by *k*-means clustering method on the set of keywords for all values of *k* between 2 and 20.

Experimental Results



SI values generated by *k*-means clustering method on the set of experts for three different keyword clustering solutions.

Experimental Results

keywords clustering	$k = 4$	$k = 6$	$k = 10$
experts clustering			
$k = 4$	0.439	0.439	0.432
$k = 7$	0.373	0.421	0.428

F-measure scores generated by *k*-means clustering method on the set of experts for $k = 4, 7$ for three different partitions of keywords ($k = 4, 6, 10$).

Experimental Results

Clusters	Keywords
1	Algorithm, Engineering, **Zoology**
2	Artificial Intelligence, Computer Science, Electrical Engineering, Computing
3	**Theory**, **Learning**, Mathematics, Electronics, Physiology, Neuroscience, Cardiology, Biochemistry, Chemistry, Biology, Molecular Biology
4	Database, Information, Software, Graphics, Botany, **Recognition**, **Privacy**, **Security**, **Parallel**
5	Medicine, Pharmacology, Ophthalmology, Toxicology, Distribute, **Pattern**
6	Data Mining, Retrieval, Energy, **Machine**, **World Wide Web**

Clustering of the set of keywords for $k = 6$.

Experimental Results

Clusters	Experts
1	▪ 27 researchers with expertise in Bioinformatics & Computational Biology, Artificial Intelligence, Data Mining and Machine Learning; ▪ All the scientists with expertise in Bioinformatics & Computational Biology; ▪ A clear sub cluster is formed by four experts all with only competence in Biochemistry. ▪ The most heterogeneous cluster.
2	▪ 9 experts with competence in Engineering, Artificial Intelligence and Computer Science.
3	▪ 12 experts with expertise in Databases and Software Engineering. ▪ Very homogeneous cluster consisting of experts all having the keyword ”Database” in her/his expertise profile.
4	▪ 5 experts: 3 with expertise in Medicine, one in Ophthalmology and one in Toxicology, Pharmacology and Molecular Biology.

Clustering of the set of experts for $k = 4$ when the keywords are partitioned in 6 clusters.

Conclusion and Future Work

- A novel semantic-aware approach for clustering of experts represented by lists of keywords has been proposed.
- The proposed approach has initially been evaluated by applying the algorithm to partition of researchers taking part in a scientific conference.
- The future aim is to pursue further enhancement and validation of our approach applying alternative clustering methods on richer expert profiles extracted from online sources.
- Our future intention is also to evaluate the scalability of the proposed approach.