Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 115–122

Semantic Explorative Evaluation of Document Clustering Algorithms

Hung Son Nguyen
Institute of Mathematics
The University of Warsaw
Banacha 2, 02-097, Warsaw Poland

Sinh Hoa Nguyen
Institute of Mathematics
The University of Warsaw
Banacha 2, 02-097, Warsaw Poland

Wojciech Świeboda
Institute of Mathematics
The University of Warsaw
Banacha 2, 02-097, Warsaw Poland

***Abstract*—In this paper, we investigate the problem of quality analysis of clustering results using semantic annotations given by experts. We propose a novel approach to construction of evaluation measure, which is based on the Minimal Description Length (MDL) principle. In fact this proposed measure, called SEE (Semantic Evaluation by Exploration), is an improvement of the existing evaluation methods such as Rand Index or Normalized Mutual Information. It fixes some of weaknesses of the original methods. We illustrate the proposed evaluation method on the freely accessible biomedical research articles from Pubmed Central (PMC). Many articles from Pubmed Central are annotated by the experts using Medical Subject Headings (MeSH) thesaurus. This paper is a part of the research on designing and developing a dialog-based semantic search engine for SONCA system[1] which is a part of the SYNAT project[2]. We compare different semantic techniques for search result clustering using the proposed measure.**

I. Introduction

CLUSTERING can be understood as an unsupervised data mining task for finding groups of points that are close to each other within the cluster and far from the rest of clusters. Intuitively, the greater the similarity (or homogeneity) within a cluster, and the greater the difference between groups, the "better" the clustering. Clustering is a widely studied data mining problem in the text domains, particularly in segmentation, classification, collaborative filtering, visualization, document organization, and semantic indexing.

It is a fundamental problem of unsupervised learning approaches that there is no generally accepted "ground truth". As clustering searches for previously unknown cluster structures in the data, it is not known a priori which clusters should be identified. This means that experimental evaluation is faced with enormous challenges. While synthetically generated data is very helpful in providing an exact comparison measure, it might not reflect the characteristics of real world data.

In recent publications, labeled data, usually used to evaluate the performance of classifiers, i.e. supervised learning algorithms, is used as a substitute [18], [6], [1]. While this provides the possibility of measuring the performance of clustering algorithms, the base assumption that clusters reflect the class structure is not necessarily valid.

[1]Search based on ONtologies and Compound Analytics
[2]Interdisciplinary System for Interactive Scientific and Scientific-Technical Information (www.synat.pl)

Some approaches therefore resort to the help of domain experts in judging the quality of the result [2], [7], [6]. When domain experts are available, which is clearly not always the case, they provide very realistic insights into the usefulness of a clustering result. Still, this insight is necessarily subjective and not reproducible by other researchers. Moreover, there is not sufficient basis for comparison, as the clusters that have not been detected are unknown to the domain expert.

There have been several suggestions for a measure of similarity between two clusterings. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data. These measures are usually tied to the type of criterion being considered in evaluating the quality of a clustering method [10].

The goal of clustering is to assign objects to subsets which are coherent internally, but are dissimilar to each other. These goals are usually explicitly formulated as *internal criteria of clustering quality*. The word "internal" highlights the fact that they are based on object similarity expressed in the original feature space. Usually it may not necessarily be clear whether modeling assumptions in the underlying model (feature space and e.g. the notion of distance between objects) are valid. Hence, one may ask to validate or evaluate a clustering in a specific application, using feedback from users or experts. When a "gold standard" clustering is provided by experts, one may compare it with the result from a clustering algorithm. This approach is an *external criterion of clustering quality*.

The problem becomes even more complicated in evaluation of text clustering with respect to semantic similarity, whose definition is not precise and highly contextual. As the number of results is typically huge, it is not possible to manually analyze the quality of different algorithms or even different runs of the same algorithm.

The remainder of this paper is structured as follows. In Section II we present some basic notions and problem statement. This is followed by an overview of external clustering evaluation methods in Section III. In Section IV we present the basic semantic evaluation techniques and propose a novel evaluation method based on exploration of expert's tags which is the fundamental contribution of this paper. After this we use the proposed evaluation method to analyze the search result clustering algorithms over the document collection publish by PubMed. An analysis of the methods result representation and their interpretability is presented in Section V, followed by some conclusions and lessons learned in Section VI.

978-1-4673-4471-5/$25.00

TABLE I. ILLUSTRATION OF HARD CLUSTERING DEFINED BY AN ALGORITHM AND BY AN EXPERT.

Doc.	Hard Cluster			Expert Cluster			
	C_1	C_2	C_3	E_1	E_2	E_3	E_4
d_1	1						1
d_2	1			1			
d_3		1					1
d_4		1				1	
d_5			1		1		
d_6			1			1	

TABLE II. ILLUSTRATION OF SOFT CLUSTERING FOUND BY AN ALGORITHM AND DEFINED BY AN EXPERT.

Doc.	Soft Cluster			Expert Tag				
	C_1	C_2	C_3	Cosmonaut	astronaut	moon	car	truck
d_1	1			1		1	1	
d_2	1				1	1		
d_3	1	1		1				
d_4		1	1				1	1
d_5		1	1				1	
d_6			1					1

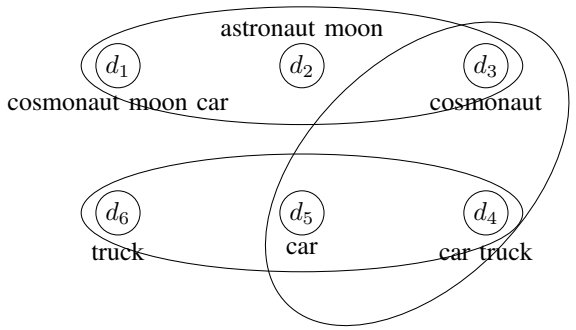

Fig. 1. Illustration of clustering from example Table II.

II. PROBLEM STATEMENT

A hard clustering algorithm is any algorithm that assigns a set of objects (e.g. documents) to disjoint groups (called clusters). A soft clustering relaxes the condition on target clusters being disjoint and allows them to overlap. Clustering evaluation measures[15], [10] proposed in the literature can be categorized as either *internal criteria of quality* or *external criteria of quality*.

An *internal criterion* is any measure of "goodness" defined in terms of object similarity. These criteria usually encompass two requirements – that of attaining high intracluster similarity of objects and high dissimilarity of objects in different clusters.

External criteria on the other hand compare a given clustering with information provided by experts. Typically in the literature it is assumed that both the clustering provided by studied algorithm and the clustering provided by experts are hard clusterings. We believe that the requirement that expert knowledge is described in terms of a hard clustering is overly restrictive. In typical applications in text mining, one faces datasets which are manually labeled by experts, but with each document being assigned a set of tags. We can think of such tags as of soft clusters. In this paper we aim to provide measures of external evaluation criteria that relax both conditions on the clustering and expert clusterings being partitions (hard clusterings).

In what follows, our focus is on clustering of documents, hence we will occasionally use terms "object" and "document" interchangeably. We stress that the notion of Semantic Explorative Evaluation (SEE) introduced in this paper is a general framework which can be used to evaluate clustering of objects of an arbitrary type.

Typically in the literature it is assumed that the input data to clustering evaluation can be described in a form similar to Table I, i.e. with exactly one valid cluster C_i and exactly one valid expert cluster E_j for each document.

We will relax this condition to allow comparison of soft clustering and a set of expert tags assigned to each document, thus allowing input data as in Table II.

III. OVERVIEW OF CLUSTERING EVALUATION METHODS

In this section we briefly review external evaluation criteria typically used in clustering evaluation. We assume that two partitions (hard clusterings) of objects are given: one by an algorithm, and another one provided by domain experts. Most external evaluation criteria can be naturally grouped in two groups:

- *Pair-counting measures*, which are defined on a 2×2 contingency matrix that summarizes similarity of pairs of objects w.r.t. both clusterings (see Table III):
  If there are k objects in the dataset, then
  $$a + b + c + d = \binom{k}{2}$$
  A typical measure that can be expressed in terms of these numbers is
  $$\textit{Rand Index} = \frac{a+d}{a+b+c+d}.$$
  However, there is a multitude of different variants of other similar measures. Pfitzner et al.[15] provide an overview of 43 measures that all fit into this scheme.
- *Information-theoretic measures* on the other hand compare distributions of $c(D)$ and $e(D)$, which denote respectively the cluster and the expert label (which induces a partition) assigned to a document D drawn randomly from the dataset. These measures can be expressed in terms of joint distribution of $\langle c(D), e(D)\rangle$, i.e. simply by counting objects belonging to each pair $\langle C_i, E_j\rangle$ as shown in Table IV. Numbers in brackets denote expected values of counts assuming independence of $c(D)$ and $e(D)$. Information-theoretic measures thus aim to to measure the degree of dependence between these two. An example such

TABLE III. ALL PAIR-COUNTING MEASURES CAN BE SUMMARIZED IN TERMS OF NUMBERS a, b, c, d IN THIS TABLE.

Pairs of documents		Same cluster? True	Same cluster? False
Same expert tag?	True	a	b
	False	c	d

TABLE IV. INFORMATION-THEORETIC MEASURES ARE DEFINED IN TERMS OF CONTINGENCY TABLE SHOWN HERE (FOLLOWING EXAMPLE IN TABLE I).

	C_1	C_2	C_3	Total
E_1	1	0	0	1
E_2	0	0	1	1
E_3	0	1	1	2
E_4	1	1	0	2
	2	2	2	

measure is *mutual information* I between $c(D)$ and $e(D)$, where

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right).$$

A measure typically used in clustering evaluation is *Normalized Mutual Information*[10], though [15] reviews 13 different measures, all defined quite similarly.

Purity is a measure occasionally used as an external evaluation criterion. While it is not strictly an information-theoretic measure, it can be also expressed in terms of table IV.

IV. SEMANTIC EVALUATION METHODS FOR SOFT CLUSTERING

We stress two limitations of measures proposed in the literature and briefly reviewed thus far:

- The first limitation, already mentioned in the previous section, is the typical assumption that both the clustering algorithm and the experts provide partitions. We will show how all measures mentioned (whether directly or indirectly) in the previous section can be naturally extended to the case of comparing a soft clustering with expert knowledge expressed in terms of multiple tag assignment. In the first part of this section we will briefly review our proposed solution to this problem, described earlier in [12].
- A more important limitation, though, is that neither of these measures described so far resemble the thought process that the expert himself would undergo if he was faced with the task of manually evaluating a clustering. In the second part of this section we will describe a novel method of semantic evaluation that addresses this issue.
  This is the fundamental contribution of this paper.
- The third problem, that we address further in the paper, is that of comparing different clusterings.
- Finally, methods mentioned so far do not allow us to compare different clusters of a single clustering.

A. Comparing set covers.

Previous works by other authors on this problem include [3] (Fuzzy Clustering Mutual Information) and [8] (comparing set covers).

In this section we consider two types of measures defined earlier separately.

TABLE V. PAIR-COUNTING MEASURES CAN BE NATURALLY DEFINED FOR SOFT COVERS IF WE SUBSTITUTE *hard membership* (SEE TABLE III) BY THE NOTION OF *similarity*.

Pairs of documents		Cluster-similar? True	Cluster-similar? False
Expert-similar?	True	a	b
	False	c	d

TABLE VI. INFORMATION-THEORETIC MEASURES CAN BE DEFINED IF WE CAN DESCRIBE THE JOINT DISTRIBUTION OF CLUSTERS AND EXPERT LABELS (SEE EXAMPLE IN TABLE II).

	C_1	C_2	C_3
Cosmonaut	0.139	0.083	0
astronaut	0.083	0	0
moon	0.139	0	0
car	0.056	0.125	0.125
truck	0	0.042	0.208

First we describe how to extend a pair-counting measure of similarity of two partitions to a measure of similarity of set covers. Pair-counting measures only pose a tiny problem. Looking at table III, we see that for soft clusterings, rows and columns are not well defined. In order to fully characterize a pair of documents $\langle d_i, d_j \rangle$, we proposed in [12] to define notions of cluster-similarity and expert-similarity for documents and base pair-counting measures on table V. This approach naturally extends any pair-counting measure, with the focus of our prior experiments on Rand Index [16]. We defined very simple notions of similarity: we considered two documents d_i, d_j θ-expert-similar, if $\frac{|e(d_i) \cap e(d_j)|}{|e(d_i) \cup e(d_j)|} \geq \theta$, and we defined θ-cluster-similarity in the same way. This approach allows us to effortlessly apply each of the 43 pair-counting measures reviewed by Pfitzner[15].

Information-theoretic measures can be extended by counting a given document in multiple cells of Table IV whenever the document is in multiple clusters and/or multiple tags are assigned to the document. If we wish to assign an overall equal weight to each document, instead of raw counts, one may further assume that the contribution of a document is inversely proportional to the number of cells that it contributes to. This has a straightforward probabilistic interpretation. The original measures, like $I(c(D), e(D))$ are defined for deterministic functions c and e and a random document D. In the proposed extension, c and e are also random variables, with $c(d)$ uniformly distributed across clusters containing document d, and $e(d)$ uniformly distributed across tags assigned to document d. Original formulas themselves, like $I(c(D), e(D))$ remain unchanged. This approach is illustrated in Table VI.

B. Semantic Explorative Evaluation

We have mentioned that the calculation of neither of the measures reviewed so far resembles human reasoning. We propose a different approach to the problem of semantic evaluation.

If an expert faced the problem of manual inspection of clustering results, he would try to explain (describe) the contents of clusters in his own words (i.e. in terms of expert tags). In essence, a cluster should be valid for an expert if the expert can briefly explain its contents. The expert would find

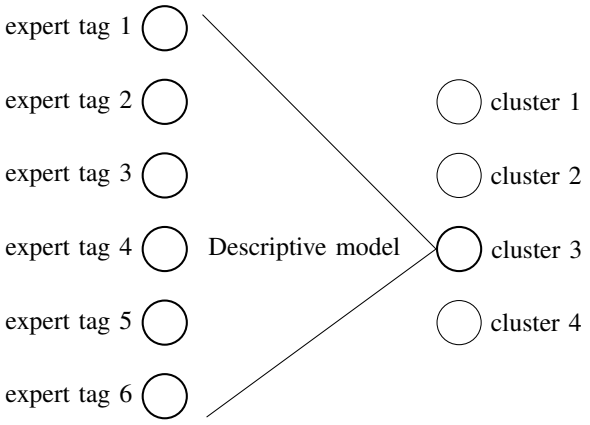

Fig. 2. For each cluster (e.g. cluster 3), we build a model describing the contents of this cluster in terms of expert tags. A measure of complexity of the resulting model thus corresponds to semantic validity of the cluster.

a set of clusters valid if he could provide a short explanation for each cluster.

In order to define a measure of semantic validity which is based on this reasoning, we need to specify three things:

- description of clusters in terms of expert concepts (i.e. a model family),
- define the length of such an explanation so that we know if it is short,
- a penalty incurred if a cluster is indescribable in terms of expert concepts,
- define the aggregate measure, so that we can evaluate a set of clusters.

We specify these three ingredients as follows:

- The explanation or description of a cluster is in essence a model of the cluster in terms of expert tags. Any classification algorithm provides such a model. The exact choice of the classifier is of secondary importance as long as the same procedure is consistently used to evaluate different clusterings. In our experiments, the classifier of choice is a decision tree with no pruning, with splits defined greedily using Gini index.
- By appealing to Minimum Description Length principle, one may then define a measure of validity of a fixed cluster as the complexity of the model describing the cluster. The measure of model complexity that we use is the average depth of the resulting tree.
- For simplicity we omit a penalty for indescribable clusters at this point, although we guarantee during tree construction that resulting leaves in descision trees contain either objects from the same decision class or objects that are indiscernible given the information about expert tags alone.
- We define the measure of validity of a clustering as the average validity of clusters.

The pseudocode of the presented idea is presented below in Algorithm 1.

Algorithm 1: SEE – Semantic Explorative Evaluation.

Input:
- $\mathbf{C} = \{C[i,j] : i = 1,\ldots,k \text{ and } j = 1,\ldots,n\}$: the document–cluster assignment matrix.
- $\mathbf{E} = \{E[i,j] : i = 1,\ldots,k \text{ and } j = 1,\ldots,m\}$: the expert–cluster assignment matrix.
- L: a decision tree construction algorithm.

Output: m: the average mean depth of decision trees describing clusters.

for $j = 1,\ldots,n$ **do**
  Construct a decision table
  $$H_j := \left[E; [C_{1,j},\ldots,C_{k,j}]^T\right]$$
  // H_j is the decision table constructed from the matrix $\mathbf{E}$ augmented with the j-th column of matrix $\mathbf{C}$ at the end as the decision variable.
  $T_j := L(H_j)$;
  // Construct the decision tree T_j by applying algorithm L on decision table H_j.
  $m_j = \text{MeanDepth}(T_j)$
end
Return $m = \frac{m_1+\ldots+m_n}{n}$;

C. Semantic Explorative Evaluation: Example

In this Section we demonstrate the proposed evaluation method (presented above in Algorithm 1) on the example introduced in Table II.

This table summarizes a small text corpus consisting of just 6 documents. Half of these documents, forming cluster C_2, concern vehicles: cars and trucks, whereas the other half concerns cosmonauts and moon: these documents form cluster C_1. Cluster C_1 is the easiest one to explain for the expert: he associates either the concept 'cosmonaut' or 'astronaut' with each document from this cluster. Document d_1 concerns a lunar rover and is an interesting "outlier" that needs to be explicitly excluded from cluster C_3 by the expert: the branch on attribute "moon" in decision tree describing cluster C_3 explicitly addresses this case.

The constructed decision trees T_1, T_2, T_3 for clusters C_1, C_2, C_3 are presented in Fig. 3, Fig. 5 and Fig. 4, respectively. According to those trees, cluster C_3 seems to be "hardest" to explain by the expert. Hence the average depth or weighted average depth of the decision tree T_3 are also higher than for T_1.

Depths of decision trees describing clusters C_1, C_2, C_3 are $1\frac{2}{3}$, 2 and $2\frac{1}{4}$, respectively. Thus SEE of the clustering equals approximately 1.97.

Doc.	Expert Tag					decision
	Cosm.	astron.	moon	car	truck	C_1
d_1	1		1	1		1
d_2		1	1			1
d_3	1					1
d_4				1	1	
d_5				1		
d_6					1	

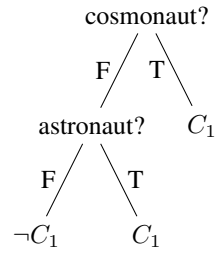

Fig. 3. The decision table H_1 (above) and the decision tree describing cluster C_1 constructed from H_1. Cluster C_1 is the easiest for the expert to explain.

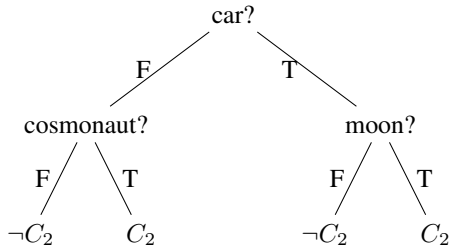

Fig. 4. Cluster C_2 is not easily describable in expert terms.

D. Randomization

The last problem we aim to address is that of comparing different clusterings. With all evaluation methods, either reviewed or introduced in this article, we face the same issue when we aim to compare different clusterings: we lack an explanation why one clustering may be better than the other one. In this section, we introduce a trick which allows us to isolate a specific sub-problem solved by a clustering algorithm, to which we can assign a measure of quality that is easily interpretable.

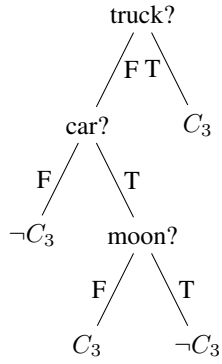

Fig. 5. Cluster C_3 does not contain document d_1, which concerns a very specific type of a car – a moon rover. This outlier forces the expert to provide a longer explanation to explicitly "remove" this object.

In what follows, we will think of a clustering algorithm as of a procedure that solves two sub-problems. For hard clustering these are:

- Determining the structure of clusters, i.e. the number of clusters and the number of documents belonging to each cluster.
- The assignment of documents to clusters, while preserving constraints on the structure.

For soft clustering, these two sub-problems are:

- Determining the structure of clusters is actually determining the number of clusters K as well as the joint (rather than the marginal) distribution of the number of documents in each cluster.
- Instead of assigning documents to clusters, an algorithm assigns documents to each of the 2^K possible partitions.

In what follows, we will focus on measuring the quality of a clustering algorithm w.r.t. the second sub-problem, while ignoring the first sub-problem. The idea is to randomize the assignment of documents to clusters while keeping the structure of clusters fixed and calculate the value of m for such randomized assignments so as to determine a meaningful "basis" or benchmark for comparison. Each measure m can thus be transformed into an m-quantile measure, which basically says how often a clustering algorithm outperforms a random assignment, while solving the second sub-problem.

V. The results of experiments

The following experiments are the continuation of our previous studies in [13], [11], [12], although in this work they merely serve as an illustration of the discussed and introduced measures.

A. Experiment Set-Up

We have applied the model-based semantic evaluation measure introduced in this paper to study clusterings induced by different document representations (lexical, semantic and structural) and using different algorithms. The document repository in our study is a subset of PubMed Central Open Access Subset[17].

TABLE VII. An example of tags assigned to the paper: "PubMed Central: The GenBank of the published literature." by Roberts R. J. ([17])

heading	subheading
Internet	
MEDLINE	economics
Periodicals as Topic	economics
Publishing	economics

The majority of documents in PubMed Central are tagged by human experts using headings and (optionally) accompanying subheadings (qualifiers) from a MeSH controlled vocabulary [20]. A single document is typically tagged by 10 to 18 heading-subheading pairs. The example of tagged document is shown in Table VII

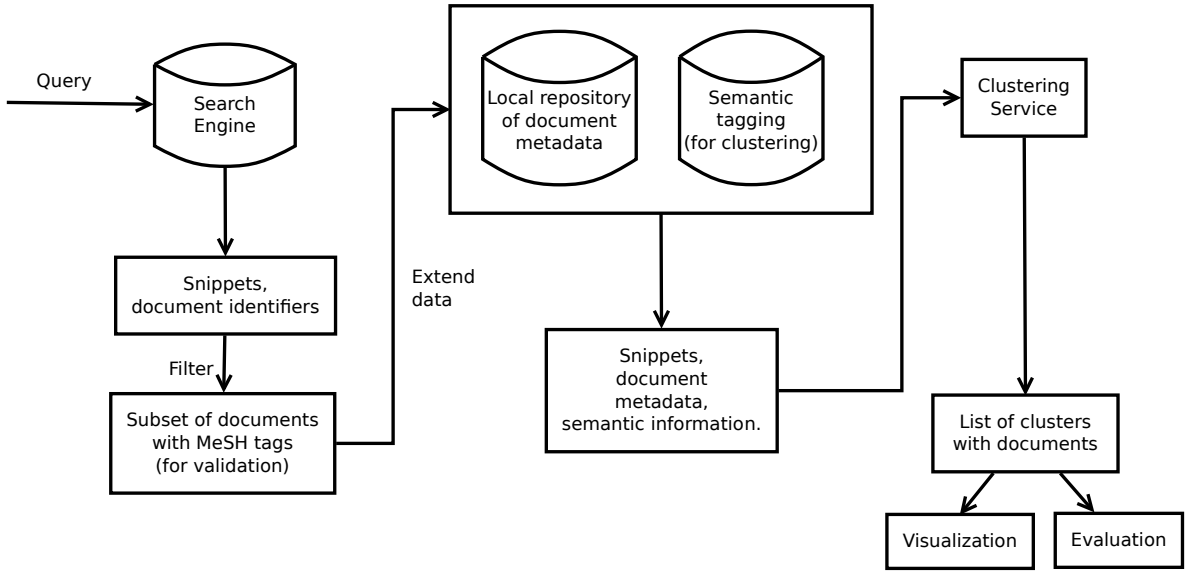

Fig. 6. Experiment diagram.

There are approximately 25000 subject headings and 83 subheadings. There is a rich structural information accompanying headings: each heading is a node of (at least one) tree and is accompanied by further information (e.g. allowable qualifiers, annotation, etc.). Currently we do not use this information, but in some experiments we use a hierarchy of qualifiers[3] by exchanging a given qualifier by its (at most two) topmost ancestors or roots.

TABLE VIII. THE NUMBERS OF POSSIBLE TAGS IN PUBMED CENTRAL OPEN ACCESS SUBSET[17]

source	**possible tags**	**expert tags assigned to example document**
headings	~ 25000	Internet, MEDLINE, Periodicals as Topic, Publishing
subheadings	83	economics
subheading roots	23	organization & administration

The choice of expert tags determines how precisely we wish to interpret expert opinion. In experiments that we describe in this paper, we interpreted subheadings as the expert tags.

The diagram of our experiments is shown in Figure 6. An experiment path (from querying to search result clustering) consists of three stages:

- Search and filter documents matching to a query. Search result is a list of *snippets* and document identifiers. Usually more than 200 documents are returned for a single query. The result set is then truncated to the top 200 most relevant (in terms of TF-IDF) documents.
- Extend representations of snippets and documents by *citations* and/or *semantically similar concepts* from MeSH ontology (these MeSH terms were automatically assigned by an algorithm[19], whereas MeSH subheadings used for evaluation were manually assigned by human experts).
- Cluster document search results.

In our experiments, we worked with three clustering algorithms: K-Means[9], Lingo[14] and Hierarchical Clustering[4].

In order to perform evaluation (and choose parameters of clustering algorithms) one needs a set of search queries that reflect actual user usage patterns. We extracted a subset of most frequent one-term queries from the daily log previously investigated by Herskovic et al. in [5] and retrieved relevant documents from PubMed Central Open Access Subset.

Roughly one fourth of these result sets was used for initial fine-tuning of parameters, whereas the remaining 71 queries were further used in evaluation.

B. Experiment results

We need to stress that we used subheadings as the source of expert tags used for semantic evaluation. There are only 83 possible subheadings in MeSH vocabulary, hence the granularity of information provided for evaluation is very limited. We have not applied pruning to resulting decision trees (the goal of algorithm Algorithm 1 is merely to provide a description, not a model for inference), and the resulting decision trees are somewhat deep, as can be seen from Figure 7.

Nevertheless, as we can see from Figure 8, the m-quantile measure is usually below 0.5 (SEE-quantile value 0.5 corresponds to a random document-to-cluster assignment). Furthermore, result sets for different queries visibly differ in how "hard" they are to cluster: m-quantile measures of different algorithms are significantly correlated. Figure 7 should not be directly interpreted in this way due to different structure of result sets corresponding to different queries (e.g. different number of documents).

[3] http://www.nlm.nih.gov/mesh/subhierarchy.html

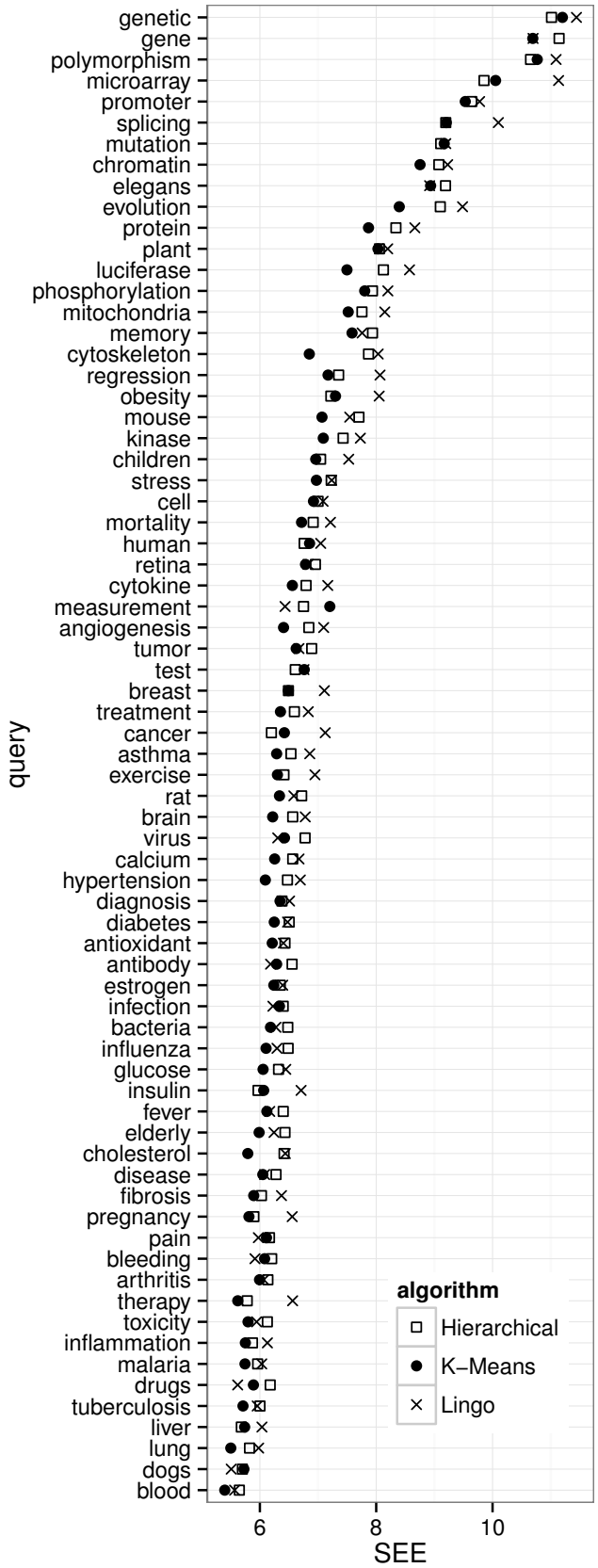

Fig. 7. Average tree depth for different result sets and clustering algorithms.

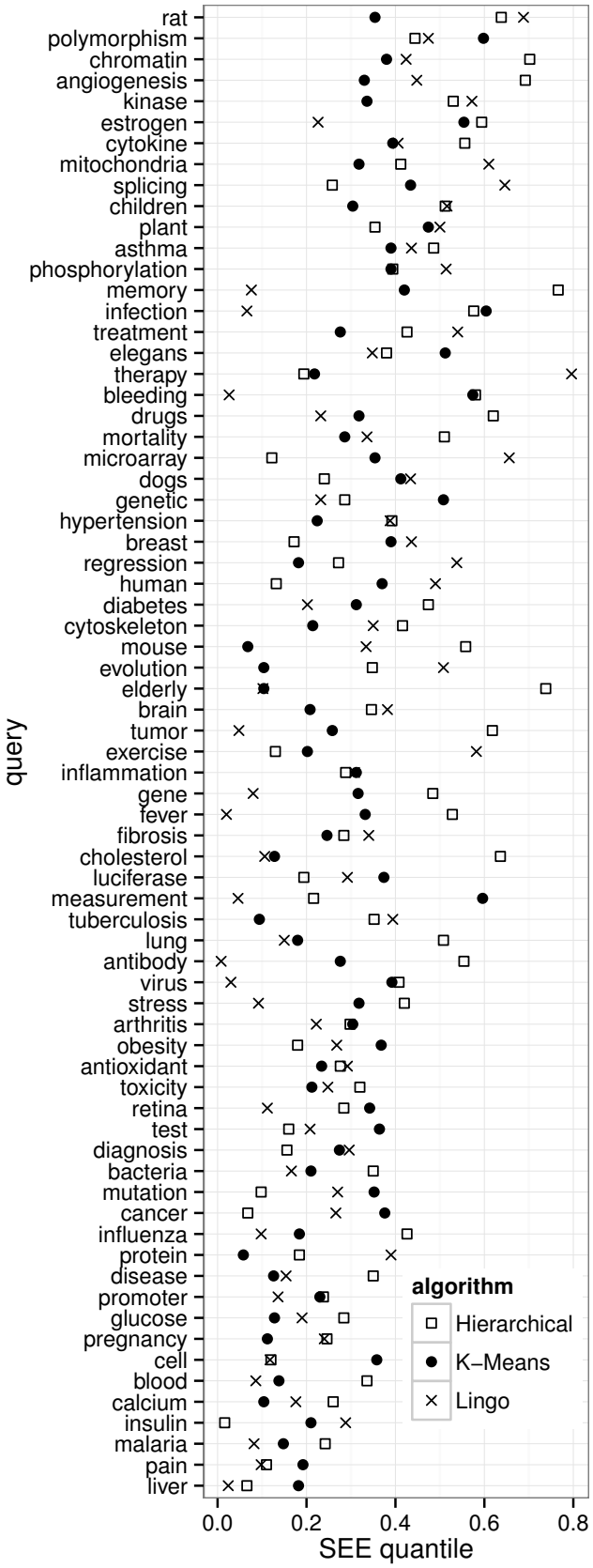

Fig. 8. SEE-quantile measure for different clusterings and queries.

VI. Conclusions and future plans

In this paper we have introduced a novel paradigm of semantic evaluation. Unlike traditional approaches, which are either measures counting pairs of objects or are variations of information theoretic approaches, our proposed procedure resembles the process of human perception, as it is based on a model describing the clustering in terms of expert knowledge. We proposed a specific implementation of this evaluation measure (i.e. a choice of the underlying model structure and optimization framework) and further demonstrated its application to online results clustering evaluation problem. We have observed that even if we only used information about MeSH subheadings assigned to documents as the source of information for evaluation, for most result sets in our experiments clusterings performed better than random assignments of documents to clusters. Furthermore, we have observed that some result sets are inherently harder to cluster than others, and the performance of analyzed clustering algorithms is usually correlated.

VII. Acknowledgements

The authors are supported by grants 2011/01/B/ST6/03867 and 2012/05/B/ST6/03215 from the Polish National Science Centre (NCN), and the grant SP/I/1/77065/10 in frame of the strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information" founded by the Polish National Centre for Research and Development (NCBiR).

References

[1] I. Assent, R. Krieger, E. Müller, and T. Seidl. Visa: visual subspace clustering analysis. *SIGKDD Explor. Newsl.*, 9(2):5–12, Dec. 2007.

[2] C. Böhm, K. Kailing, H.-P. Kriegel, and P. Kröger. Density connected clustering with local subspace preferences. In *ICDM*, pages 27–34. IEEE Computer Society, 2004.

[3] T. Cao, H. Do, D. Hong, and T. Quan. Fuzzy named entity-based document clustering. In *Proc. of the 17th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2008)*, pages 2028–2034, 2008.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[5] J. R. Herskovic, L. Y. Tanaka, W. Hersh, and E. V. Bernstam. A day in the life of pubmed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association*, pages 212–220, 2007.

[6] H.-P. Kriegel, P. Kroger, M. Renz, and S. Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 250–257, Washington, DC, USA, 2005. IEEE Computer Society.

[7] P. Kröger, H.-P. Kriegel, and K. Kailing. Density-connected subspace clustering for high-dimensional data. In M. W. Berry, U. Dayal, C. Kamath, and D. B. Skillicorn, editors, *SDM*. SIAM, 2004.

[8] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 11:033015, March 2009.

[9] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* 2007.

[11] S. H. Nguyen, W. Świeboda, and G. Jaśkiewicz. Extended document representation for search result clustering. In R. Bembenik, Ł. Skonieczny, H. Rybiński, and M. Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 390 of *Studies in Computational Intelligence*, pages 77–95. Springer-Verlag New York, 2012.

[12] S. H. Nguyen, W. Świeboda, and G. Jaśkiewicz. Semantic evaluation of search result clustering methods. In R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 393–414. Springer, 2013.

[13] S. H. Nguyen, W. Świeboda, G. Jaśkiewicz, and H. S. Nguyen. Enhancing search results clustering with semantic indexing. In *SoICT 2012*, pages 71–80, 2012.

[14] S. Osinski, J. Stefanowski, and D. Weiss. Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Systems*, pages 359–368, 2004.

[15] D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowl. Inf. Syst.*, 19(3):361–394, May 2009.

[16] W. M. Rand. Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.*, 66(336):846–850, 1971.

[17] R. J. Roberts. PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):381–382, Jan. 2001.

[18] K. Sequeira and M. Zaki. SCHISM: A new approach for interesting subspace mining. In *Proceedings of the fourth IEEE conference on Data Mining*, pages 186–193. IEEE Computer Society, 2004.

[19] M. Szczuka, A. Janusz, and K. Herba. Semantic clustering of scientific articles with use of dbpedia knowledge base. In R. Bembenik, Ł. Skonieczny, H. Rybiński, and M. Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, pages 61–76. Springer-Verlag New York, 2012.

[20] United States National Library of Medicine. Introduction to MeSH – 2011. Available online, 2011.