Comparative cluster labelling involving external text sources

Ferenc Kruzslicz
PhD, Associate Professor
University of Pécs
E-mail: kruzslic@ktk.pte.hu

Balázs Kovács
Assistant Lecturer
University of Pécs
E-mail: kovacsb@ktk.pte.hu

Miklós Hornyák*
Assistant Lecturer
University of Pécs
E-mail: hornyakm@ktk.pte.hu

Giving clear, straightforward names to individual result groups of clustering data is most important in making research usable. This is especially so when clustering is the real outcome of the analysis and not just a tool for data preparation. In this case, the underlying concept of the cluster itself makes the result meaningful and useful. However, a cluster comes alive only in the investigator's mind since it can be defined or described in words. Our method introduced in this paper aims to facilitate and partly automate this verbal characterisation process. The external text database is joined to the objects of the clustering that adds new, previously unused features to the data set. Clusters are described by labels produced by text mining analytics. The validity of clustering can be characterised by the shape of the final word cloud.

KEYWORDS:
Cluster quality.
External index.
Comparative word cloud.

DOI: 10.20311/stat2017.K21.en101

* Supported by the ÚNKP-17-3-IV.-PTE-193 New National Excellence Program of the Ministry of Human Capacities.

Clustering techniques assign data objects (entities) to (sub)groups not defined in advance based on the similarity of their properties. They attracted attention long before the beginning of big data trends. Whilst the concept itself appeared very early in the field of cultural anthropology (*Driver–Kroeber* [1932]), its use became routine only from the 1960's due to its computational requirements. Thanks to the advances in computing, basic multivariate mathematical-statistical algorithms are now available in all standard software packages. However, a major limitation to the applicability of clustering methods remains: they provide results within a reasonable time only for data sets with a relatively small number of elements. In addition, the quality of the result is not easy to assess and depends on the type of clustering. Thus, among the 5Vs of big data (volume, variety, velocity, veracity, and value), the volume of data and the velocity of data generation represent the greatest challenge for clustering. At the same time, the processing of non-numerical data and uncertain information are no longer obstacles to such algorithms. Roughly in parallel with the types of clustering tasks becoming clear, the *Hungarian Statistical Review* published the first two articles on cluster analysis methods by *József Csicsman* and *Péter Futó* in 1979 (*Csicsman* [1979], *Futó* [1979]). They proved to be forward looking in many ways over the years. First, they showed that clustering tasks are equivalent to finding the quasi components of a (hyper)graph. Second, they demonstrated that their method is also suitable for clustering textual data. The cluster analysis method that they defined from an information science point of view is in fact equivalent to the document-clustering problem defined in the field of text mining.

Time changed not only the terminology but also the purpose of clustering. The primary application areas were typing, model fitting, estimation, and hypothesis testing based on groups. All of this has been supplemented by pattern discovery in large data sets, compression of the data set, and hypothesis generation, partially as a response to the needs posed by big data. In the case of information retrieval systems, for example, large amounts of textual data can be semantically compressed by news clustering, so that the responses to user search queries will be faster and more accurate. Our method builds upon a similar document clustering technique, or more precisely, on a labelling technique.

Clustering is primarily not the result of data mining methodologies used to extract patterns from large data sets, but a data preparation and data-cleansing tool. These initial steps are crucial for the outcome of the analysis, and so cluster analysis is performed in several stages in an iterative manner. In order to achieve better data quality, one must be able to compare different cluster structures to each other to find the most suitable one for the task. The paper of *Füstös–Meszéna–S.-né Mosolygó* [1977] published in *Szigma* was the first article in Hungarian literature dealing with cluster validation.

Cluster validity can be determined primarily in two ways. If we are constrained to use the same data during the clustering and validation phases, then we call this unsupervised validation. If, however, we compare the clustering to an already known external structure (e.g. existing classes), then we measure the degree of fit using a supervised method. The cluster validation method proposed in this paper is a mixture of the two. Our validation method utilises external textual information not used in the clustering phase, but this information does not have any structure required by supervised methods. In a big data context, it is usually easy to assign textual attributes to objects of clustering based on the author or the subject of the text.

Clustering validity can be characterised not only by indices but also by other methods. Examples for this are the fact that a test algorithm can be performed on the clusters, or that there exists a visualisation technique which confirms the existence of the clusters. However, although each alternative cluster validation method could ultimately be converted to indices, we would lose a good deal of additional information by doing so. In addition, it is difficult to judge whether a particular index value indicates adequate clustering. Indices may be helpful at most in choosing between two clustering performed by the same type of method on the same data set. Partly, this is the source of the belief that clustering is more art than science. Our proposed cluster labelling method does not quantify the quality of the clustering; instead it applies a new visualisation technique which essentially tells us how accurately individual clusters can be textually described and distinguished from each other. This is a special version of the technique that became popular with the proliferation of infographics, a version of the word cloud diagram based on *Drew Conway*’s idea (*Conway* [2011]).

In the following parts of the study, we first review the existing cluster labelling techniques, cluster validation methods and word cloud variations, after which we present our text-based labelling method as a combination of the above. The use of the method is illustrated by an example utilising an online data source.

1. Document clustering

Document clustering is no more than the application of traditional clustering techniques to text files. Clustering can be done without developing special algorithms working on textual data if we can find a suitable numerical representation of documents. Among them, the simplest one – which we also use – is vector space representation that describes a document *d* as the vector $\mathbf{d} = \left(x_1, x_2, \ldots, x_m\right)$ in the domain of the space of possible words. The set of observed documents is termed the corpus. The x_i element of a document vector is an index related to the i^{th} word within the dictionary of the words of the corpus. This index is the term frequency of the given word in the case of the most commonly used bag of words model. It should be noted that this involves a loss of information since we lose the word order, but, in return, all the usual clustering operations can be performed in such a vector space. The interpretation of the results of calculations performed in this space can cause some difficulties, such as in the case of document vectors resulting from average calculations that may consist of non-integer numbers. In such a case, we can either use only clustering techniques adequate for nominal and ordinal data types, or assign approximate documents to the resulting un-interpreted document vectors. The method of finding the median string (*Kruzslicz* [1999]) can be directly integrated into prototype-based clustering algorithms. There are several possible clustering approaches but we deal only with the partitioning methods which assign objects to disjoint groups. We do not exclude the applicability of any other (e.g. fuzzy, hierarchical) method; we only assume that the results of these algorithms are always converted into partitions.

The documents of the corpus are often converted into vectors not in their original form, and they go through various preparatory transformation steps. Tokenisation is the process in which words and punctuation marks are separated from each other. The task of stemming is to replace the different forms of word occurrence with a common stem. This can be done by a simple suffix-stripping algorithm in a language-independent manner or by applying appropriate lemmatisation rules after language detection. Besides these syntax-based solutions, there are also semantic attempts that take into consideration also synonymous words (not necessarily included in the corpus) as a replacement. During stop word removal, those words which belong to the most common but least meaningful words of the particular language, such as pronouns, are deleted from the document. These transformations cause information loss also, and their main objective is to reduce the dimensionality of the vector space. It is generally accepted that the above steps are only necessary to reduce the computational time and storage requirements of the modelling. If sufficiently many documents were available, text mining models themselves would carry out these tasks automatically for us, but for the time being, it takes a lot more time and is less accurate. Data preparation could be extended with other expert knowledge in addition to the linguistic knowledge if the corpus belonged to a well-defined topic. The word “statistics”, for example, would be qualified probably as stop word in the corpus of the articles of the *Hungarian Statistical Review*, and “Gaussian distribution” would be a compositional phraseme.

In addition to the bag of words model, many other document representation models are also known. The state transition matrix, as a description of word sequentiality, for example, can be used to partially preserve the word order, or the document vector of the standard bag of words model can be supplemented with the part of speech of the words, or semantic elements could be incorporated into the model through phrase structure trees. External lexical sources, such as Wikipedia and WordNet, can also be involved in the preparation process after we determined the language of the individual parts of the document. However, only a few are supported by standard software packages. The main reason for this is that complex data structures greatly increase the computational requirements of the algorithms.

A bag of words representation makes it possible to use not only the traditional algorithms at the clustering phase, since we can even rely on the well-known methods during validation. Assuming that documents are similar to each other if they contain nearly the same words and roughly in the same proportion, we can derive the similarity measure used in clustering from the cosine distance measure. The cosine distance between two documents (**d₁** and **d₂**) can be determined by the following equation, where the numerator contains the scalar product of the two vectors, and $\|\mathbf{d}\|$ refers to the length of the vector:

$$\cos\left(\mathbf{d}_1,\ \mathbf{d}_2\right) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\left(\|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|\right)}.$$

Given a cluster function, we can use any cohesion (intra-cluster compactness) index, (inter-cluster) separation index or a combination of these to determine cluster validity. Using a distance-based cluster function, a smaller cohesion value and a bigger separation value means better clustering. The supervised, unsupervised and relative indices of clustering are surveyed in more detail in the articles of e.g. *Legány–Juhász–Babos* [2006] and *Rendón et al*. [2011] dealing with this subject.

2. Document labelling

Labelling textual data is defined as the assignment *T* which assigns a set of words (i.e. a label) $T\left(\mathbf{d}\right) = \left\{t_1, t_2, \ldots, t_k\right\}$ with a variable number of elements by document to each document **d**, in a way that this set describes the content of the document and distinguishes it from the other documents best. The former definition changes in the case of clustering only in respect of assigning labels to a set of similar documents $\mathbf{D} = \left\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\right\}$ instead of a single document **d**.

The words of the label set are not constrained, and we do not assume any connection between them. If the data structure of the labels were a tree structure instead of a set structure, we would be dealing with hierarchical labelling (category system), or, if it were a directed graph, we would call it taxonomy. In the special case of hierarchical labelling, we must also take into consideration the parent-child relationships of the label categories when naming the clusters automatically. *Mao et al*. [2012] defined two criteria for both sibling and parent-child relationships of labels to construct a correct labelling. The first and the fourth criteria for the sibling relationships are a reformulation of the previously mentioned criteria for word sets:

> – According to the first principle, the labels of a cluster should be representative of important terms in that cluster.
>
> – Pursuant to the fourth principle, a label is preferred the more for a sibling category the less often it occurs as a frequent term in other sibling clusters.

However, we also need two other criteria to define the correctness of a label hierarchy:

> – In conformity with the second principle, the label of the parent category should also be a common term to each of its child clusters.
>
> – In keeping with the third principle, the label of the parent category must be more general than the labels of its child categories.

The usage of ontologies[1] requires additional naming validity principles.

We can utilise the structure of the clustering to properly, and from other clusters distinctively, label an entire cluster. Moreover, we can use the intermediate results of the clustering algorithm as *Mei–Shen–Zhai* [2007] did with non-negative matrix factorisation and with latent Dirichlet allocation. Early labelling methods used, for instance, information inherent in the merge (or splitting) tree of the hierarchical clustering to find the terms that distinguish the vertically related clusters from each other. An easier solution is to merge the texts of the documents of the cluster into a new document and to consider the label assigned to this document the label of the cluster. We can do this because we assume that each cluster is already made up of similar documents, and this merged document can be considered the centroid of the cluster. However, it is not necessary to merge the documents. It is also a viable option to

[1] Ontology: a formal explicit specification of a shared conceptualisation (*Gruber* [1993]).

label each document before clustering, and after it, we combine the label sets of the documents belonging to the same cluster (for example, as the union or the intersection of the labels). Document labelling is a type of compression method in which, according to *Manning–Raghavan–Schütze* [2008], the most descriptive phrases can be found basically by two methods: cluster-internal labelling or differential (or comparative) cluster labelling. In respect of cluster-internal labelling, we use only the documents of the cluster to find the correct labels. This method carries the risk that the common terms of a cluster may not be specific only to that given cluster. Therefore, it is strongly recommended to remove stop words, both general and topic specific words, from the documents before performing cluster-internal labelling on them. In respect of differential cluster labelling, we also use the content of other clusters, and so this method is more efficient than the previous one. There are several principles for selecting the labels in this way. We can use keyword or entity (i.e. person, location, etc.) detection algorithms. The output of text extraction and summarisation methods can also be converted to label sets. Note, however, that, if we know the set of possible labels beforehand, we are facing a classification task and not a clustering task. Therefore, the determination of the set of keywords to be used as labels is also a part of document labelling, and this can be done most easily by weighting the words of the corpus.

We can measure the fitness of such a word to a document (or to a group of documents) by a combination of two indices. The term frequency d_{ij} tells us the number of times the j^{th} term of the corpus is used in the i^{th} document. The document frequency f_j tells us the number of documents the j^{th} term is used in. Term frequency is therefore a cluster-internal index, but document frequency is a comparative index. A particular term is the more characteristic of a document, the more it is used in it, i.e. the greater its word frequency is, and the lower its document frequency is. The product of the inverse of the document frequency, and the term frequency results in the *tf-idf* index (term frequency-inverse document frequency) which has the following formula:

$$tf\text{-}idf\left(t_j, \mathbf{d}_i\right) = d_{ij} \cdot \log\left(n/f_j\right),$$

where n is the number of documents.

Terms with a high *tf-idf* value are the most adequate for labelling. In many cases, however, the best labels are not even mentioned in the document, which is common mainly in hierarchical category systems. In the case of scientific papers, for example, the higher-level topic-specific terms appear only in the topic description of the journal that published the paper. The article itself contains only more specific terms belonging to a subtopic. Therefore, it is recommended to involve external big data sources into labelling (e.g. databases with a category system such as Wikipedia

(*Carmel–Roitman–Zwerdling* [2009]) or CiteSeerX (*Kashireddy–Gauch–Billah* [2013])).

Similarly to cluster validity, the correctness of the labelling can be measured in two ways. Supervised indices assume that the correct labels of the documents are known, and they measure the degree of match between the predicted and correct labels (*Treeratpituk–Callan* [2006]). Examples are the accuracy, purity, recall, precision, normalised mutual information, rand, and other indices, which can be found in nearly every data mining handbook (*Aggarwal–Zhai* [2012], *Tikk* [2007]). If the documents in question are not associated with relevant labels to compare the results with, then external text sources with a category system, or text mining methods can be useful, assuming that the foreign language equivalents of the documents, or their counterparts with similar contents are labelled. These methods, however, are usually only significant from the developmental and testing perspective of the algorithms. Real world problems often demand unsupervised label validity methods, but it is very difficult to automate this process without the involvement of external resources. Under experimental conditions, it is common to validate the labelling manually by human experts (*Pantel–Ravichandran* [2004], *Maqbool–Babri* [2006], *Geraci et al.* [2006]), but this is, of course, hardly an objective measure.

3. Text databases

We showed in the previous chapter that it is recommended to involve external corpora and other linguistic databases into labelling. In the following sections, we describe the two most commonly used ones of these data sources and present some of the interesting results and solutions of the related literature.

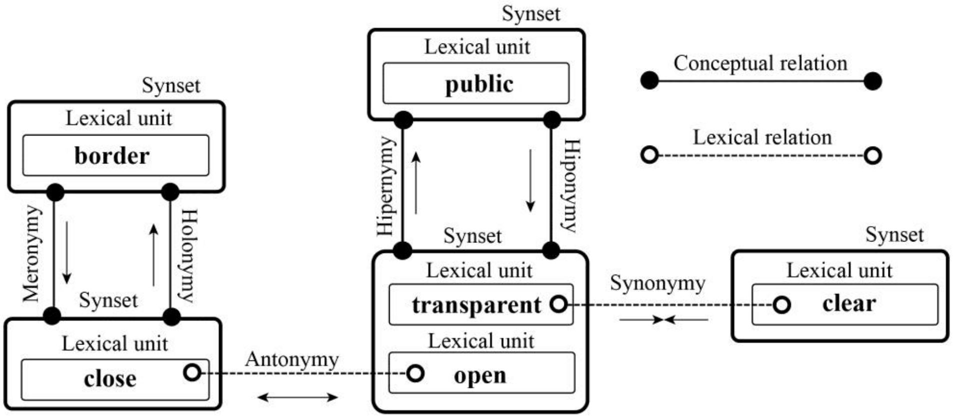
3.1. WordNet

WordNet is one of the largest and most widely used English lexical databases, which is developed by researchers at Princeton University.[2] It contains about 150 thousand lexical entries organised into synsets as shown in Figure 1. These lexical units enable WordNet to represent the vertical and horizontal relationships of a word as a tree. Relationships between words with similar and opposite meanings in the database are considered synonymous and antonymous, respectively, whilst the

[2] wordnet.princeton.edu

most important hierarchical relationships are generalisation-specialisation (hypernymous and hyponymous) and part-whole (meronymous and holonymous) relationships.

Figure 1. An example of synset relations in the WordNet database



Source: Authors' own creation.

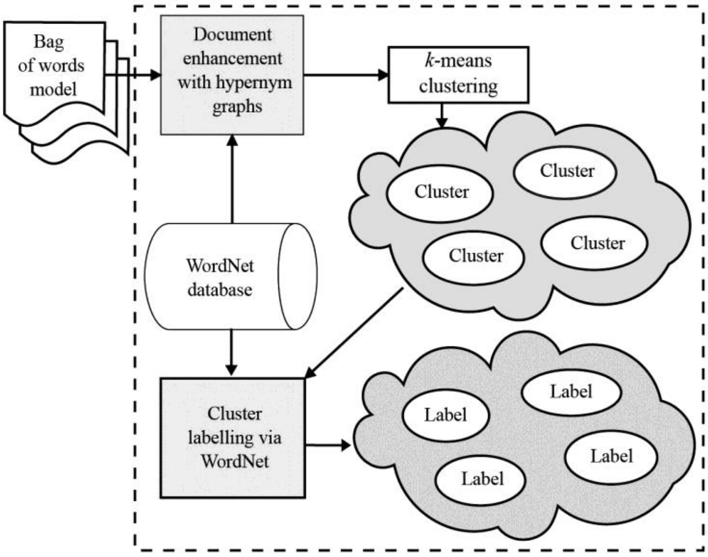
These relationships between lexical units allow us to use the WordNet system to measure semantic similarity, since the closer to each other two terms are, the more similar is the structure of their relationship, and the more common are the lexical units shared. One of the first similarity measures of this type (*wup*) was defined by *Wu–Palmer* [1994]. Here we give a few examples of the *wup* measure based on WordNet:

```
wup (statistics, mathematics)  = 0.900,
wup (statistics, economics)    = 0.762,
wup (statistics, distribution) = 0.444,
wup (statistics, probability)  = 0.267.
```

Shenoy–Shet–Acharya [2012] surveyed the semantic similarity-based clustering techniques and measures, and they pointed out that the set of terms produced by techniques utilising the relationships between lexical units could be too heterogeneous, and that the algorithms should be able to handle this situation. The data structure of WordNet is not only suitable for the creation of a new kind of cluster functions, but it can also be utilised during data preparation. *Gharib–Fouad–Aref* [2010] used the WordNet database to improve the quality of stemming. They substituted the stems obtained by conventional stemming algorithms with their hypernym in the WordNet database. In this way, they were able to reduce the number of dimensions in their model and increase the efficiency of traditional clustering algorithms on text data at the same time. Using WordNet, we can enrich our corpus with an additional external semantic structure, so it is a useful tool when we face labelling, especially cluster-internal labelling, since the structure of relationships adds information to the texts.

Bouras–Tsogkas [2012] developed the *W*-kmeans algorithm shown in Figure 2, which uses the WordNet database to enhance simultaneously both clustering and labelling performance. Their clustering method is not based on the basic bag of words model, instead it utilises hypernym graphs belonging to the top 20% of the most frequent terms. Examples: *statistics* → (*datum, data_point*) → *information* → (*cognition, knowledge, noesis*) → *psychological_feature* (*abstraction, abstract_entity*) → *entity.* They combined the resulting trees and weighted the words in them according to their frequency and position inside the tree. After this, they added the new terms of the combined tree with the greatest weights to the document. Thereafter, they performed a traditional *k*-means clustering on the documents represented by the semantically enriched content. They used the same method at the labelling phase, but instead of using all the documents, they used the top 10% of the terms of the cluster and assigned the top five terms of the combined hypernym graph by weight to the cluster as label.

Figure 2. The sketch of the W-kmeans algorithm



Source: Adapted from *Bouras–Tsogkas* [2012].

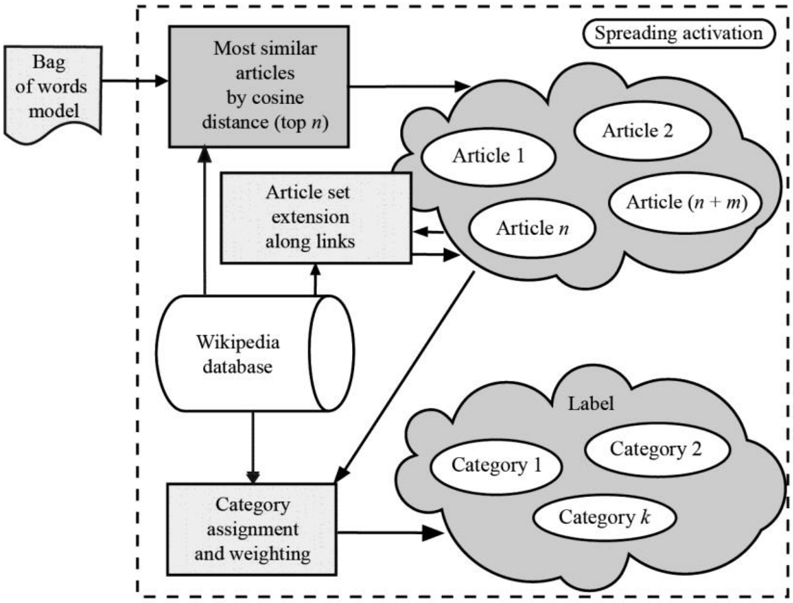
3.2. Wikipedia

Wikipedia,[3] the collaborative encyclopaedia editable by anyone, was launched in 2001 by *Jimmy Wales*, and it is probably used and known by more people than WordNet. Wikipedia with its more than 5 million English articles has grown into one of the largest text corpora in the world. It is particularly suitable for developing machine translation and cross-lingual text mining methods, since an article is translated into multiple languages. Another important feature of Wikipedia is that it is organised into categories what makes it a good external data source for cluster labelling. It is important to note that the category system of Wikipedia serves primarily indexing purposes, and thus it does not strictly have a tree-like structure. That is, a subcategory may belong to several main categories. The keywords of Wikipedia contents can be organised into an ontology which relate the keywords of the documents to each other or even to other external, formal ontologies (e.g. DBpedia, Semantic MediaWiki), through hyperlinks. Wikipedia as an ontology is sufficiently comprehensive, mostly of high quality, adapts rapidly to novelty, and is easily understood.

Syed–Finin–Joshi [2008] constructed a graph based on Wikipedia whose nodes correspond to the titles of the articles (concepts), and the directed edges between the nodes correspond to the links between the articles. They defined several spreading activation labelling heuristics on this graph. Each of these methods has a certain number of starting points in their initial set of activated nodes, and additional nodes will be activated in each iteration along the joined edges. The spread of the activation can be controlled through various restrictive conditions (e.g. distance or degree limits). Their first method started by finding the most similar Wikipedia articles to the documents based on the cosine measure, and the associated categories were gathered. The categories were weighted based on the number of their occurrences or their aggregate cosine similarity, and the most suitable labels for the analysed documents were chosen from the top-weighted categories. Their second method differed from the first in that the gathered categories served as the initial set for several iterations of the activation spreading algorithm applied on the graph of Wikipedia categories. Finally, they performed the regular weighting-based labelling on the set of categories expanded as mentioned previously. Figure 3 shows a third method that utilised the information inherent in the links within the articles. After they determined the most similar Wikipedia articles to the document, they applied the spreading algorithm several times on the links within the articles to add more articles to this set. They prevented the spread along links pointing to irrelevant sources by deleting edges from the graph of Wikipedia articles whose cosine similarity falls below a certain value (0.4). The labels were determined by the above method, but they used the expanded set of articles.

[3] wikipedia.org

Figure 3. The outline of the spreading activation algorithm



Source: Adapted from *Syed–Finin–Joshi* [2008].

4. Building category systems

We mentioned formerly that Wikipedia is an article database with a category system whose categories were constructed manually by authors and editors, and they update them regularly to reflect new events and people. The spreading activation algorithm can be extended to other databases that contain interconnected documents assigned to categories. If there are no links between the documents, this method is no longer viable, but the categorised corpora can still be used to assess the quality of the labelling algorithms. This can be done by first assigning labels to the analysed documents, then finding those texts in the database, which are the most similar to each document and, finally, comparing the documents' labels with these texts' categories. Basically, every web portal with a menu structure can be considered such a database. Typical examples are news websites with topics (domestic, sport, economy, etc.). The *Reuters* news agency even made available a corpus[4] of 10 788 articles (contain-

[4] http://about.reuters.com/researchandstandards/corpus/

ing 1.3 million words) for research purposes in 2000, whose documents are classified into 90 categories.

Web catalogue services, the first generation of search engines, were built upon classification as well. Yahoo![5] is the most famous of these, whose category system, called Yahoo Directory, was also constructed manually. Though the service was shut down in 2014, its last state (with 1.8 million items) is still available for research purposes. Recently, the Open Directory Project took over this role, which is one of the largest general-purpose, manually constructed online category systems for web pages with its more than 4 million items.[6] There are also certain topics or areas with unique category systems. The ACM CCS (Association for Computing Machinery Computing Classification System), for example, is a six-level ontology, whose latest, 2012 version has 2 113 categories for scientific publications relating to computer science and information processing.[7] Cluster analysis occupies multiple positions in this hierarchy:

Computing methodologies
  ∟ Machine learning
    ∟ Learning paradigms
      ∟ Unsupervised learning
        ∟ Cluster analysis

Mathematics of computing
  ∟ Probability and statistics
    ∟ Statistical paradigms
      ∟ Cluster analysis

The manual construction and the update of a category system demand significant resources, and so there is a strong urge to automate these activities. The construction of taxonomies is nothing other than finding the best hierarchical clustering of a document set based on a certain similarity function and then assigning the best labels to these clusters. Typically, we cannot rely on other external data sources whilst building category systems. This, however, is of less significance in the case of ontologies for scientific publications, because the authors must precisely declare the subject and the contribution of the paper to the literature by selecting keyword categories.

Kashireddy–Gauch–Billah [2013] developed a method for the automatic classification of the scientific articles of the CiteSeerX corpus[8], which is capable of labelling every new category of the expanding ontology. Initially, only 2.6% of the 2 million documents of the CiteSeerX corpus were labelled based on the top three levels of the CCS ontology. They used a k-NN classification method to classify the rest of the corpus into the pre-existing categories. The disadvantage of this method is that the number of elements in each category amounts to tens of thousands, which makes it inconvenient for browsing for users. This could, however, make the ontolo-

[5] yahoo.com

[6] dmoztools.net

[7] http://www.acm.org/about/class/

[8] citeseerx.ist.psu.edu

gy more accurate by a further breakdown of the third-level categories, and this improved the thematic retrievability of the articles. They divided the publications of the large categories into *k* clusters using a partitioning algorithm, and the clusters were named based on an automatic labelling algorithm. They selected a hundred documents from each cluster randomly and determined the part of speech (POS) of each word in those documents in order to find the right label for them. Cluster label candidates were selected from among the nouns only. They ranked the potential labels based on the *tf-idf* measure as well as two of their own indices (*delta-tf* and *tf-stdev*). The *delta-tf* index is calculated by subtracting the average term frequency of the other clusters from the term frequency of the given cluster. The *tf-stdev* weighting scheme takes the inter-cluster variability of the term into account by multiplying the term frequency of the given cluster by the standard deviation. They measured the efficiency of their method on the documents with known category labels and found that the *tf-stdev* weighting scheme gave the best results. This test was carried out by selecting the top three terms for each weighting scheme and comparing them to the manually labelled category names.

5. Measuring the quality of labelling

Corpora with category systems are significant for labelling, as they make it possible to measure the efficiency of labelling algorithms. If the real labels are known, the accuracy of a particular labelling can be observed. The paper of *Kashireddy–Gauch–Billah* [2013] illustrates how hard it is to achieve entirely accurate labelling. For example, if a manual label is not present in any of the documents, it is impossible to judge its adequacy without the involvement of external data sources. Consequently, first we have to define label matching. The *k* parameter in the definition means that only the top *k* items of the ranked list of labels are considered. Every usual classification performance measure is applicable to this top-*k* method, and we conventionally put the @K notion after the name of the index in this case for clarity.

Precision (*Precision@K*) is a measure comparing the number of matches for a label to the number of times the labelling algorithm assigned the same label to any document. Recall (*Recall@K*), on the other hand, is a measure comparing the number of matches for a label to the number of times the same label was manually assigned to any document. The F_1 measure is calculated as the harmonic mean of these two indices. The results of *Kashireddy–Gauch–Billah* [2013] for the CiteSeerX corpus, using the *tf-stdev* weighting scheme and the “at least one of the top three” definitions of label matching, are: *Precision*@3 = 0.47, *Recall*@3 = 0.56, and therefore F_1@3 = 0.55.

Carmel–Roitman–Zwerdling [2009] relabelled a random sample of documents from the Open Directory Project[9] (an open-content directory of World Wide Web links) and the 20 Newsgroups[10] (20 thousand news articles in twenty newsgroups) corpora automatically. They also used WordNet to measure the performance of their labelling. Their definition of label matching requires the label to match either the name of the category or any synonym of the corresponding lexical unit in the WordNet database in any inflected form. The *Match@K* measure used by the authors indicates the percentage of clusters where the label matches at least one of the top *k* items of the weighted list of label candidates. They observed that, without the involvement of external data sources, approximately 15% of the documents' text does not contain the proper category label. This means that it is impossible to improve the quality of the automatic labelling based on the vocabulary of the documents. On the other hand, long top-lists are impractical and the *k* parameter should not be much greater than 5. Lower *k* values, however, resulted in *Match@K* values lower than 0.7 and 0.5 in their two corpora, respectively. Finally, they examined the remaining 85% of the documents and found that feature selection methods rarely evaluate manual labels as appropriate. This shows that manual labels, in many cases, are irrelevant according to the statistical methods.

In order to improve the labelling, the authors involved Wikipedia in the process as an external data source. They constructed a search query from the most important words of the clusters and analysed the metadata of the Wikipedia pages returned by the query. More precisely, terms in the title and in the category of the articles were added to the set of terms extracted from their texts, and then the authors weighted and ranked the entire set of documents. With this method, the value of the *Match@K* index has been improved by 10-40% depending on the value of the *k* parameter. In case the value of *k* was 5, the value of the *Match@*5 index was above 0.85 in both corpora.

The researchers used another index called *Mean Reciprocal Rank@K*. The reciprocal rank of a cluster is defined as the reciprocal of the position of the first matching label in the top-*k* list of the cluster label candidates – or zero if there was no match. In the case of multiple matches, only the first one is taken into account. *MRR@K* is basically the average of the reciprocal ranks of the clusters.

Mao et al. [2012] performed hierarchical labelling on documents sampled from Yahoo Answers[11] and the main categories of the Wikipedia corpus. They distinguished exact and partial label match depending on whether the automatically assigned label matches the labels of both the category and its parent category or merely one of them. Synonyms of the labels were also considered as matches. They calcu-

[9] http://dmoztools.net/

[10] http://qwone.com/~jason/20Newsgroups/

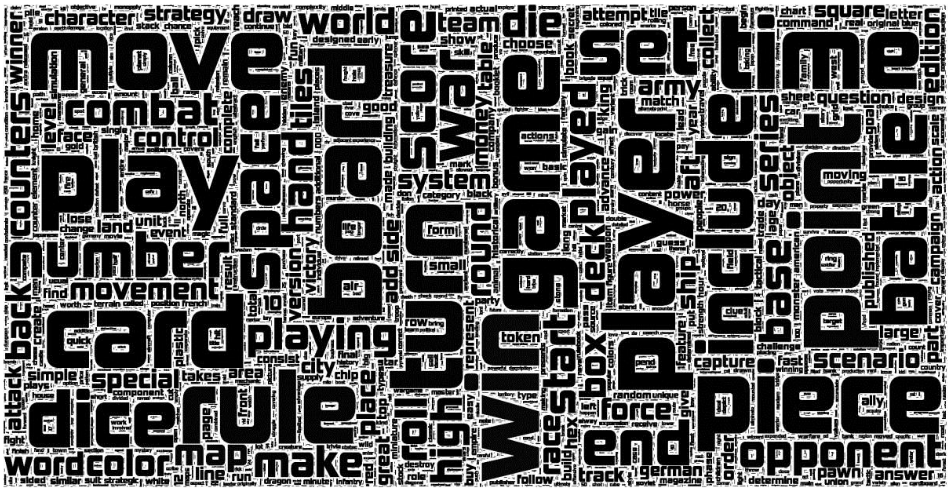
[11] https://answers.yahoo.com/

lated the indices mentioned earlier to both corpora using 1, 3, and 5 as the value for *k*. They found that the value of *Match@K* is increasing and the value of *Precision@K* is declining with the increase of *k*. For comparison, the best value for the exact *Match@*5 index was 0.448 in both corpora, whilst the best partial index values were 0.867 and 0.673 in the Yahoo! and the Wikipedia corpora, respectively.

6. Word cloud diagrams

Figure 4 shows an example of a word cloud, a visualisation technique that represents data points as the texts of their labels, and the other properties of these texts, such as position, orientation, font, size, colour, etc., are based on values associated with the data points.

Figure 4. A word cloud overview of the descriptions from the BoardGameGeek board gaming database



Source: Authors' own creation.

Milgram–Jodelet [1976] were the first who used this visualisation technique. They wrote the names of the main tourist attractions of Paris on the map of the city in a way that the text size was proportional to the popularity of the place. Zeitgeist, the first computer program, which generated word clouds automatically, was released in 1997, and it visualised the search queries entered on a website as a diagram embeddable into webpages. *Jim Flanagan*, the author of the Perl script took advantage of the feature of the HTML language regarding the simplicity to adjust the colour and the font size of the words according to the number of their occurrences. The usage of word clouds became particularly widespread from the 2000s on, when the first online word cloud web services, among which the most famous is Wordle[12], were released. The opinions of researchers were divided as to whether or not early word cloud visualisations were simply plotting frequency table data, which provided a difficult and inaccurate way to obtain information. Therefore, many research papers investigated how to make the word cloud visualisation technique more effective.

Bateman–Gutwin–Nacenta [2008] studied how font, size, colour, intensity, text length, orientation, position, and the area and resolution of the diagram affect readability. *Lohmann–Ziegler–Tetzlaff* [2009] investigated how word order, layout and grouping help to interpret word clouds better, and they also tracked the eye movement of the users and observed which part of the diagram the users remembered the most. The word cloud is not only a design tool but also a user navigation tool and a fundamental tool of infographics – attributable to the fact that it supports the following four activities quite well:

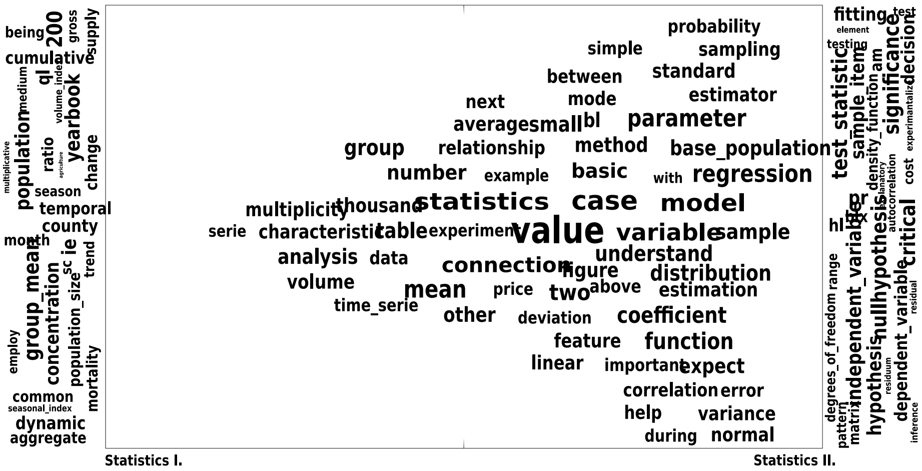
> – *Searching*: purposefully looking for something (or a lack of it), rapid identification of alternative results.
>
> – *Browsing*: discovery without strict purpose arousing a user's interest in one or more areas.
>
> – *Impression*: gaining a general impression of the specified underlying content and relevant concepts during reading.
>
> – *Pattern recognition*: anticipation and recognition of deeper relationships between the analysed concepts based on their co-occurrence.

As word clouds gained popularity, an increasing number of publications presented new functionalities (e.g. shape filling, text rotation). More and more complex algorithms were created to fill optimally the available space on the chart. *Burch et al.* [2013] developed an arranging algorithm that produces a "prefix word cloud" from the terms of the same root. *Jänicke et al.* [2015] upgraded the conventional pie chart into the so-called TagPie by incorporating word clouds into it. This is also a good example of parallel tag clouds which visualise several corpora at the same time. *Collins–Viégas–Wattenberg* [2009] examined the usability of different implementations of parallel tag clouds as interactive user interfaces. *Carpendale et al.* [2010] developed the so-called SparkCloud based on the sparkline visualisation technique, which is useful in recognising trends and changes in textual data. *Diakopoulos et al.* [2015] created Compare Cloud that displays terms frequently co-occurring with a certain keyword in two corpora and shows that of which corpus each term is more characteristic.

[12] wordle.net

A list of label candidates, consisting words and their weights, is perfectly suitable for word cloud creation. The resulting word cloud gives a more detailed characterisation of the cluster as if only the first *k* elements of the list were used. Compare clouds can visualise the structure of two of the cluster label candidate sets at a time, whilst parallel tag clouds can display all at once. From this point, we will use an improved version of Conway's comparative word cloud as shown in Figure 5. The middle section of the chart shows the terms that can be found in both documents. The size (or occupied territory) of the words is determined according to the number of combined occurrences. Since a written word is a two-dimensional object, its font size is proportional to the square root of its frequency. If the two analysed corpora comprise multiple documents, it is recommended that more sophisticated weighting methods be used (e.g. *tf-idf*). The horizontal position of words within this area shows the relative frequency of occurrences in the two corpora. There are terms scaled similarly but pooled together at both ends of the chart, each representing one of the corpora. The font colour serves solely the separation and readability of the words. The font sizes in the middle and at the two ends of the chart are not comparable with each other. The common words occur more frequently than single words, and so the font size of the word clouds at the two ends must be enlarged for readability. It is also recommended that the number of words appearing on the chart be limited for clarity reasons. The usual data preparation steps (tokenising, stemming, stop word removal) were carried out before generating the figure, and only the most frequent of the common and unique words are displayed.

Figure 5. Top 30 comparative word cloud – comparison of Herman–Rédey [2006] and Pintér–Rappai [2006] textbooks



Source: Authors' own creation.

After discussing the structure of comparative word clouds, we can read from Figure 5 that, despite the different authors, the terminology of the two textbooks is not significantly different. The books are based on each other, and the second volume largely uses the terms of the first. The unique words at the two ends of the chart show that the subjects of the textbooks are clearly distinguished from each other.

7. Cluster evaluation through comparative labelling

Please note that if Figure 5 did not show a comparison between two documents but rather two document clusters, then their degree of separation would be represented as the essence of the unique words and the "butterfly shape" of the common words concentrated near the ends of the chart. Cluster cohesion would materialise in the proportion and position of large and small words. In this way, we can visually confirm to what extent objects of the two clusters are grouped around certain concepts. Our cluster validation method is based precisely on this observation.

The process of labelling-based cluster validation is as follows:

1. Objects are clustered in an arbitrary way.

2. If the objects have not previously been associated with texts, then appropriate related documents must be looked for and assigned to them.

3. A weighted list of label candidates must be constructed from each document set associated with a cluster.

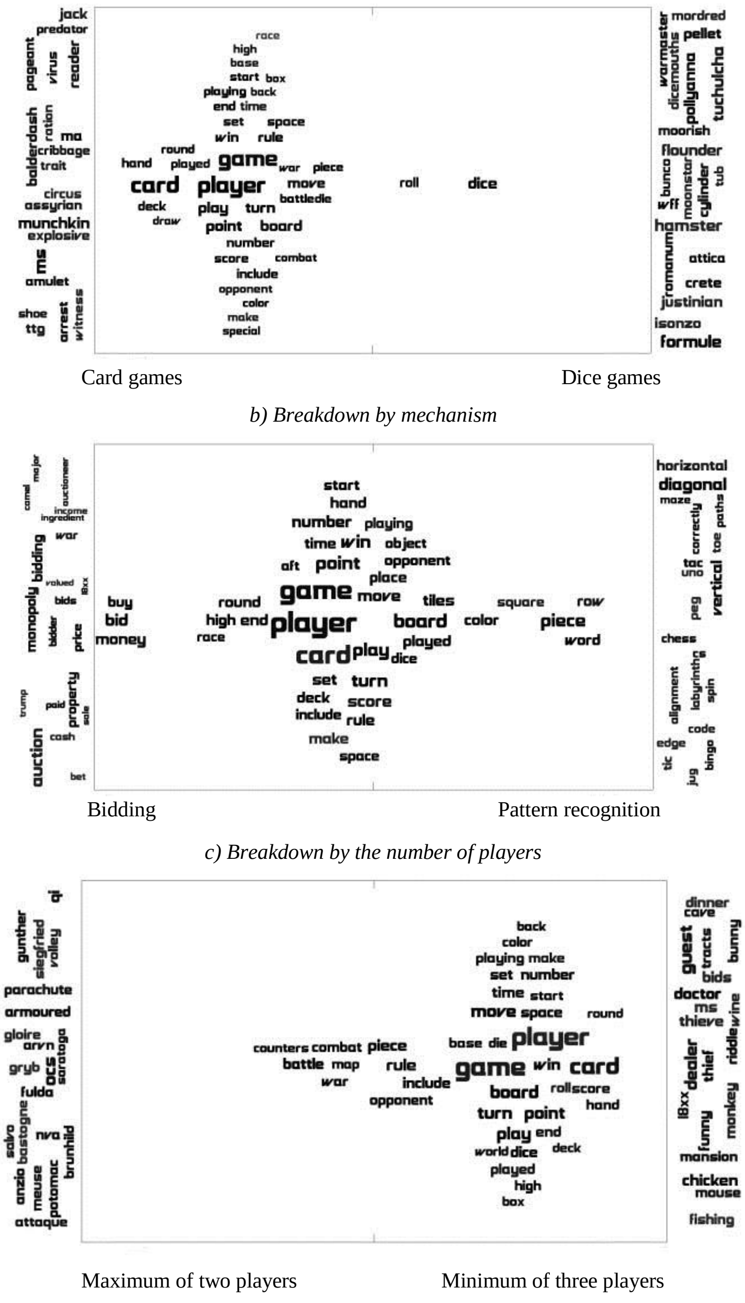
4. A comparative word cloud is generated from each pair of cluster label sets.

5. Based on these comparative word clouds, it can be evaluated how well defined and distinguishable the clusters are.

We wished to demonstrate the operation of the method by means of an empirical example and to do this, we had to find a database with objects with many attributes, to which documents could easily and clearly be assigned. We did not choose the BoardGameGeek database[13] of board games only because the games can be grouped based on many different attributes, but also because the average person has sufficient background knowledge of this subject to evaluate the results. We randomly selected 10 thousand board games from the database but did not perform clustering to keep our results reproducible; instead, we divided them into groups using predefined queries. We tested our method by assessing whether the comparative world clouds reflected our assumptions that were used to formulate the queries.

[13] http://boardgamegeek.com/

Figure 6. Comparative word clouds of board game groups

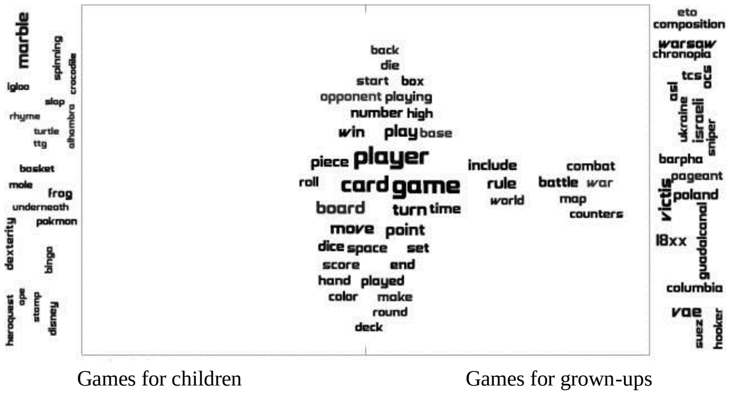


a) Breakdown by category

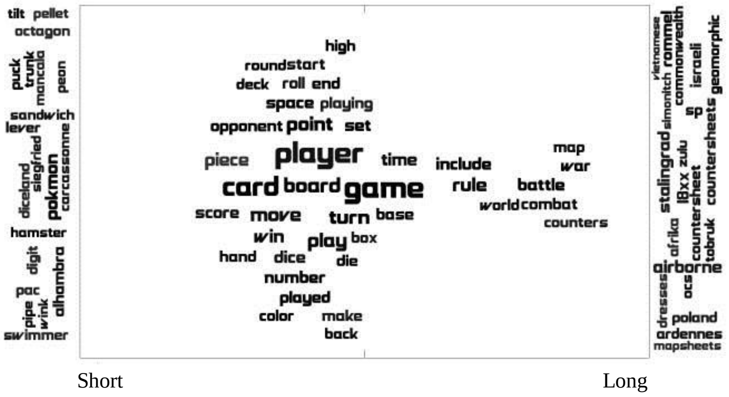
b) Breakdown by mechanism

c) Breakdown by the number of players

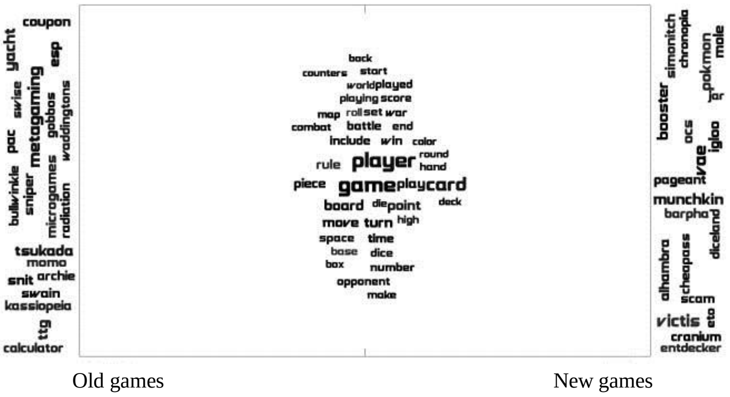
d) Breakdown by minimum age



e) Breakdown by playing time



f) Breakdown by publication year



Note. Frequency tables for the source of the charts are available on the online Annex (www.ksh.hu/statszemle).

Source: Authors' own creation.

The most important attributes used in our queries are: name, year of publication, minimum and maximum number of players, game time and recommended lower and upper age limit, game designer and publisher, category of the game type and mechanics, required language skills, rating and rank of the game. We assigned their short descriptions to the games as textual documents.

The word cloud generated after standard data preparation is large, and its frequent words are the special, not too surprising stop words of the topic (play, player, board, figure, card, move, etc.). Figure 6 shows the comparative word clouds of several pairs of label sets using only the top twenty most frequent common and unique words. The diagrams also contain the topic-specific stop words, since the position of these words also indicates the relative size of the sets. Excluding, however, the terms typical of all board games, we could achieve a clearer picture of the differences of the sets.

a) In respect of *breakdown by category*, we expect that the existing category labels describe their elements appropriately and accurately. The actual labels of the two selected categories, card and dice, were correctly placed at the respective ends of the chart. These terms are also large sized, which means that the two sets of documents are well separated and coherent. Note that the figure also indicates what actions are associated with them: cards are meant to be drawn and dice to be rolled. The positions of common words (game, player) are biased towards the left side of the chart meaning that the sizes of the two sets are significantly different. Indeed, there are many more card games (1 644) than dice games (595) in the sample selected from the database. The unique terms (the names of subcategories within the categories) may not be familiar to outsiders. These are identifying terms rather than generalising ones, and so a further breakdown of the two sets is not necessary.

b) In respect of *breakdown by mechanism*, a similar conclusion can be drawn, and only the number of elements in the two sets is more balanced: 405 bidding games and 318 pattern recognition games are included. The other difference is that many more words are needed to label the set, since the word "mechanism" either does not appear in the descriptions or appears very seldom. The characteristics of bidding games are the use of money, the purchase, the offer, and the auction; whilst pattern recognition requires to find words, colours, tiles or directions. Unique words are more general than in the previous case, and so they are more suitable for characterising the particular mechanism. In the case of large databases, it may be necessary to further break down the groups. Therefore, grouping games by similar mechanism produces a better breakdown than doing so by category.

c) In respect of *breakdown by the number of players*, it is clear that there are many more multiplayer games in the sample as well as in reality. The unique words of games for a maximum of two players are identified again, and the unique words of games for a minimum of three are rather general ones. The terminology of the two groups does not differ significantly, but war-/battle-/combat-theme games are more common among games for two, and the partner more often is called an opponent. If there are multiple players, however, the order (round) and the first player (start) must be defined as well. The two sets, however, cannot be clearly distinguished from each other without knowing the number of players.

d) At first sight, *breakdown by minimum age* is surprisingly similar to the breakdown by the number of players. By choosing a ten-year age limit, it was expected that games for small children do not require many skills, and writing, reading and calculation are not requisites. By contrast, the word cloud showed other characteristics. As the unique words of the games for children indicate, children's games are distinguished from others mainly by their themes (animals, cartoon figures). The common words show that the role of chance is greater as the term "roll" is somewhat shifted towards the side of the children's games. War-themed games are relatively far from the world of children's games and are characteristic of games for grown-ups. The unique words of games for grown-ups are rather general, which means that these games require geographical and historical knowledge. The two sets cannot be well differentiated, because a game could easily be adapted to the other group by changing the appearance of the game.

e) In respect of *breakdown by playing time*, unique words are of identifying nature, and so we cannot give a general characterisation of the reasons why some games require more than sixty minutes whilst others do not. The common words show only that war games and games requiring maps tend to last longer. Furthermore, together with the previous two charts, we can conclude that breakdowns by the number of players, minimum age, and playing time are much more interesting combined than in themselves. Together they suggest that games for grown-ups, in contrast to children's games, last longer, and children's games are rather multiplayer games, whilst duels are typical for grown-ups' games. This could mean that we tried to characterise a group in all three cases, which has a more complex definition than those attributes.

f) The intention behind the *breakdown by publication year* was that we would examine whether a trend in the evolution of games is visible. As can be seen, we divided the sample into two roughly equal parts using the date 1990 as the threshold value. It was found that the unique words identify primarily the names of the games, whilst the terminologies of the old and new games are almost entirely the same. This type of breakdown is, therefore, completely pointless; the game is never over.

8. Summary

We demonstrated a visual cluster validation method and its backgrounds in this paper. The pairwise comparison of cluster label candidates can be performed in each case when textual documents or sets of documents can be assigned to the clustered objects. The novelty of our method, on the one hand, lies in a semi-supervised technique that helps to identify the meaning and evaluate the quality of the clusters by utilising external textual data sources in a way never studied before. On the other hand, the structure of the comparative word cloud used for the analysis is also an algorithmic and, in terms of its appearance, upgraded version of a previous implementation. Every small step forward may be required as cluster validation is still not highly developed and is a less-used area of cluster analysis, although it should be a mandatory part of any such analysis. Our proposed method is not nearly an exact method, but it is an advancement compared with the completely subjective cluster validation and characterisation. The subjective elements make it difficult to learn its application, but the examples presented show that it is a very useful tool.

The technique shown is quite general and can be customised and upgraded as needed. First, it may be worthwhile to consider which of the sophisticated weighting methods would produce better charts instead of the simple frequency measure. Another possibility to improve the method would be the topic-specific stop word removal as mentioned in the empirical example. The current implementation of the comparative chart assigns data only to the horizontal position and font size. It could also be examined how to display more attributes on the diagram using font colour, vertical position, orientation, contrast and other parameters. An interactive version of the comparative word cloud might also be helpful during the manual labelling process to choose the best candidates. The biggest challenge is the implementation of multi-way comparisons that would make it possible to analyse more than simply two clusters at a time.

The crucial point of the method is the opportunity for easily joining good quality unstructured data to the structured data in future systems, and so we expect that big data trends will point in the direction of storing data in a less and less structured form, and that technologies will become available to retrieve easily the structured data needed for analysis.

References

AGGARWAL, C. C. – ZHAI, C. X. [2012]: *Mining Text Data*. Springer. Cham. http://dx.doi.org/10.1007/978-1-4614-3223-4

BATEMAN, S. – GUTWIN, C. – NACENTA, M. [2008]: *Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections.* Proceedings of the Nineteenth ACM (Association for Computing Machinery) Conference on Hypertext and Hypermedia. ACM. New York. pp. 193–202. http://dx.doi.org/10.1145/1379092.1379130

BOURAS, C. – TSOGKAS, V. [2012]: A clustering technique for news articles using WordNet. *Knowledge-Based Systems.* Vol. 36. pp. 115–128. http://dx.doi.org/10.1016/j.knosys.2012.06.015

BURCH, M. – LOHMANN, S. – POMPE, D. – WEISKOPF, D. [2013]: *Prefix Tag Clouds*. Proceedings of the 17th International Conference on Information Visualisation. Institute of Electrical and Electronics Engineers Computer Society. Washington, D.C. pp. 45–50. http://dx.doi.org/10.1109/IV.2013.5

CARMEL, D. – ROITMAN, H. – ZWERDLING, N. [2009]: *Enhancing Cluster Labeling Using Wikipedia*. Proceedings of the 32nd International ACM SIGIR (Association for Computing Machinery Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval. ACM. New York. pp. 139–146. http://dx.doi.org/10.1145/1571941.1571967

CARPENDALE, S. – KARLSON, A. K. – LEE, B. – RICHE, N. H. [2010]: SparkClouds: visualizing trends in tag clouds. *Visualization and Computer Graphics*. Vol. 16. No. 6. pp. 1182–1189. http://dx.doi.org/10.1109/TVCG.2010.194

COLLINS, C. – VIÉGAS, F. B. – WATTENBERG, M. [2009]: *Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora*. Proceedings of the Symposium on Visual Analytics Science and Technology. Institute of Electrical and Electronics Engineers. New York. pp. 91–98. http://dx.doi.org/10.1109/VAST.2009.5333443

CONWAY, D. [2011]: Building a better word cloud. *R-bloggers*. 27 January. http://drewconway.com/zia/2013/3/26/building-a-better-word-cloud

CSICSMAN, J. [1979]: A klaszter-elemzés módszerei és alkalmazási lehetőségei a statisztikában. *Statisztikai Szemle*. Vol. 57. No. 2. pp. 137–145.

DIAKOPOULOS, N. – ELGESEM, D. – SALWAY, A. – ZHANG, A. – HOFLAND, K. [2015]: *Compare Clouds: Visualizing Text Corpora to Compare Media Frames*. Proceedings of the IUI (Intelligence User Interfaces) Workshop on Visual Text Analytics. http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Visualizing-Text-Corpora-to-Compare-Media-Frames_final.pdf

DRIVER, H. E. – KROEBER, A. L. [1932]: *Quantitative Expression of Cultural Relationships*. University of California Press. Berkeley.

FÜSTÖS, L. – MESZÉNA, GY. – S.-NÉ MOSOLYGÓ, N. [1977]: Cluster analízis: fogalmak és módszerek. *Szigma.* Vol. X. No. 3. pp. 111–148.

FUTÓ, P. [1979]: Hipergráf modellen alapuló klaszter-elemzés és alkalmazása. *Statisztikai Szemle*. Vol. 57. No. 2. pp. 130–136.

GERACI, F. – PELLEGRINI, M. – MAGGINI, M. – SEBASTIANI, F. [2006]: Cluster generation and cluster labelling for web snippets. *Lecture Notes in Computer Science*. Vol. 4209. pp. 25–36. http://dx.doi.org/10.1007/11880561_3

GHARIB, T. F. – FOUAD, M. M. – AREF, M. M. [2010]: Fuzzy document clustering approach using WordNet lexical categories. In: *Elleithy, K.* (ed.): *Advanced Techniques in Computing Sciences and Software Engineering*. Springer. Cham. http://dx.doi.org/10.1007/978-90-481-3660-5_31

GRUBER, T. R. [1993]: A translation approach to portable ontology specifications. *Knowledge Acquisition.* Vol. 5. Issue 2. pp. 199–220. http://dx.doi.org/10.1006/knac.1993.1008

HERMAN, S. – RÉDEY, K. [2005]: *Statisztika I*. Pécsi Tudományegyetem. Pécs.

JÄNICKE, S. – BLUMENSTEIN, J. – RÜCKER, M. – ZECKZER, D. – SCHEUERMANN, G. [2015]: *Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies*. Göttingen Dialogue on Digital Humanities. Göttingen. http://www.etrap.eu/wp-content/uploads/2015/05/Stefan_Jaenicke.pdf

KASHIREDDY, S. D. – GAUCH, S. – BILLAH, S. M. [2013]: *Automatic Class Labeling for CiteSeerX*. Proceedings of the IEEE (Institute of Electrical and Electronics Engineers)/WIC (Web Intelligence Consortium)/ACM (Association for Computing Machinery) International Conference on Web Intelligence. IEEE, WIC, ACM. New York. pp. 241–245. http://dx.doi.org/10.1109/WI-IAT.2013.35

KRUZSLICZ, F. [1999]: Improved greedy algorithm for computing approximate median strings. *Acta Cybernetica*. Vol. 14. Issue 2. pp. 331–339.

LEGÁNY, CS. – JUHÁSZ, S. – BABOS, A. [2006]: *Cluster Validity Measurement Techniques*. Proceedings of the 5th WSEAS (World Scientific and Engineering Academy and Society) International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. WSEAS. Madrid. pp. 388–393.

LOHMANN, S. – ZIEGLER, J. – TETZLAFF, L. [2009]: *Comparison of Tag Cloud Layouts: Task-related Performance and Visual Exploration*. Proceedings of the 12th IFIP (International Federation for Information Processing) TC13 (Technical Committee on Human–Computer Interaction) International Conference on Human-Computer Interaction: Part I. Springer-Verlag. Berlin, Heidelberg. pp. 392–404. http://dx.doi.org/10.1007/978-3-642-03655-2_43

MANNING, C. D. – RAGHAVAN, P. – SCHÜTZE, H. [2008]: *Introduction to Information Retrieval*. Cambridge University Press. Cambridge. http://dx.doi.org/10.1017/CBO9780511809071

MAO, X. – MING, Z. – ZHA, Z. – CHUA, T. – YAN, H. – LI, X. [2012]: *Automatic Labeling Hierarchical Topics.* Proceedings of the 21st ACM (Association for Computing Machinery) International Conference on Information and Knowledge Management. ACM. New York. pp. 2383–2386. http://dx.doi.org/10.1145/2396761.2398646

MAQBOOL, O. – BABRI, H. A. [2006]: Automated software clustering: an insight using cluster labels. *The Journal of Systems and Software*. Vol. 79. Issue 11. pp. 1632–1648. http://dx.doi.org/10.1016/j.jss.2006.03.013

MEI, Q. – SHEN, X. – ZHAI, C. [2007]: *Automatic Labeling of Multinomial Topic Models*. Proceedings of the 13th ACM SIGKDD (Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery and Data Mining. ACM. New York. pp. 490–499. http://dx.doi.org/10.1145/1281192.1281246

MILGRAM, S. – JODELET, D. [1976]: Psychological maps of Paris. In: *Proshansky, H. M. – Ittelson, W. H. – Rivlin, L. G.* (eds.): *Environmental Psychology*. 2nd Edition. Holt, Rinehart & Winston. New York. pp. 104–124.

PANTEL, P. – RAVICHANDRAN, D. [2004]: *Automatically Labeling Semantic Classes.* Main Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. Boston. pp. 321–328.

PINTÉR, J. – RAPPAI, G. [2005]: *Statisztika II. A következtetéses statisztika alapjai*. Pécsi Tudományegyetem. Pécs.

RENDÓN, E. – ABUNDEZ, I. – ARIZMENDI, A. – QUIROZ, E. M. [2011]: Internal versus external cluster validation indexes. *International Journal of Computers and Communications*. Vol. 5. Issue 1. pp. 27–34.

SHENOY, M. K. – SHET, K. C. – ACHARYA, D. U. [2012]: A new similarity measure for taxonomy based on edge counting. *International Journal of Web & Semantic Technology*. Vol. 3. Issue 4. pp. 23–30. http://dx.doi.org/10.5121/ijwest.2012.3403

SYED, Z. S. – FININ, T. – JOSHI, A. [2008]: *Wikipedia as an Ontology for Describing Documents.* Proceedings of the Second International Conference on Weblogs and Social Media. AAAI Press. Palo Alto. pp. 136–144. http://ebiquity.umbc.edu/paper/html/id/383/Wikipedia-as-an-Ontology-for-Describing-Documents

TIKK, D. [2007]: *Szövegbányászat*. TypoTex Kiadó. Budapest.

TREERATPITUK, P. – CALLAN, J. [2006]: *Automatically Labeling Hierarchical Clusters*. Proceedings of the 2006 International Conference on Digital Government Research. Digital Government Society of North America. San Diego. pp. 167–176. http://dx.doi.org/10.1145/1146598.1146650

TSOGKAS, V. – BOURAS, C. [2012]: A clustering technique for news articles using WordNet. *Knowledge-Based Systems*. Vol. 36. December. pp. 115–128. https://doi.org/10.1016/j.knosys.2012.06.015

WU, Z. – PALMER, M. [1994]: *Verb Semantics and Lexical Selection*. Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics. Association for Computational Linguistics. Stroudsburg. pp. 133–138. http://dx.doi.org/10.3115/ 981732.981751