Article

Visual Exploration of Medical Records

Úrsula Torres-Parejo [1,†,‡*], Jesús R. Campaña [2,‡*], M. Amparo Vila [2,‡*] and Miguel Delgado [2,‡*]

[1] Department of Statistics and Operational Research, University of Granada; ursula@ugr.es
[2] Department of Computer Science and Artificial Intelligence, University of Granada; (jesuscg,vila,mdelgado)@decsai.ugr.es
* Corresponding authors
† Current address: Avda. Fuente Nueva s/n 18071, Granada, Spain
‡ These authors contributed equally to this work.

Abstract: Medical records contain many terms which are difficult to process. Our aim in this study is to allow the visual exploration of the information in medical databases where the texts presents a large number of syntactic variations and abbreviations, through an interface which facilitates content identification, navigation and information retrieval. We propose the use of multi-term tag clouds as content representation tools and as assistants for the browsing and querying tasks. The tag cloud generation is achieved through a novelty mathematical method that allows related terms to remain grouped together within the tags To evaluate this proposal, we have used a database with 24,481 records. 23 expert users in the medical field were tasked to complete a survey to evaluate the generated tag clouds properties and we obtained a precision of 0.990, a recall of 0.870 and a F1score of 0.904 in the evaluation of the tag cloud as an information retrieval tool. The main contribution of this approach is that we automatically generate a visual interface over the text capable of capturing the semantics of the information and facilitating access to medical records.

Keywords: knowledge representation; electronic health records; health information systems; content identification; visual interface

Citation: Torres-Parejo, U.; Campaña, J.R.; Vila, M.A; Delgado, M; Visual Exploration of Medical Records. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

In the medical field multiple data are collected every day. In order to be useful, data must be processed, which is complex task [1]. In emergency management, surgical interventions, human resources and all other hospital areas it is necessary to extract knowledge from data that would help the general management of the hospital, patients' care and decision making.

As far as textual information it is concerned, this is not an easy task, since a text may present a large number of syntactic variations or even mistakes. In addition, this information is usually inserted by different people who use different writing patterns. The number of semantically relevant entities grows constantly and rapidly as new scientific discoveries are made [2,3], so it is necessary to have intelligent systems that process data semantically, as well as syntactically, which add structure, and facilitate the formulation of semantic queries about textual attributes. Furthermore the search for clinical information is becoming a critical technique for rapid and effective access to patients' information [4], so it is essential to have a simple interface which helps in the query formulation and does not require prior knowledge for its use.

To address these problems, we propose a method to process the textual information contained in these databases preserving the semantics and offer a tag cloud as graphical interface that represents the content of the information allowing the identification of entities and their relationships. This graphical interface also allows the user to browse the contents of the database and query in an intuitive way.

Tag clouds seem to be a good alternative due to the familiarity that most of the users have with them as internet navigation tools and their ability to represent information content [5]. They are already used in the medical field, for biomedical text summarization as described in [6], and in other studies with a similar purpose to ours but with a totally different generation method, resulting in a mono-term tag cloud [7].

doi:10.20944/preprints202107.0624.v1

The mono-term tag clouds, mostly used in other approaches, produce some important drawbacks regarding the ability to identify contents [8–10]. Think for example terms commonly used in medicine such as "central line" or "posterior chamber", the meaning of these terms is not inferred by the union of the meanings of the words that compose them. So, for a correct identification of the information content, these terms must remain within the tag cloud. Another drawback of the mono-term tag clouds is that they do not allow the identification of the relationships between terms, which also affects semantics.

On the other hand, one of the main disadvantages attributed to tag cloud visualizations is that they do not have a standardized generation method and that they lack an underlying mathematical model, which is especially useful since it allows the definition of the operations performed in the database [10,11].

In addition, whilst the associations between entities usually involve only two entities, in the medical field relationships may involve more than two (for example "bone lesion biopsy"), which involves complex associations [12].

To overcome all these challenges we propose an automatically generated tag cloud. This tag cloud would allow the user to browse and query the database and facilitate content identification and the relationships between the concepts due to the use of multi-term tags. It is defined mathematically, and is obtained through a standardized method.

1.1. Background

1.1.1. Information Extraction in the Clinical Domain

Technologies to process text in natural language for the extraction of information were introduced into the medical field more than two decades ago [13], and were developed and used for many systems in different applications [14,15].

Information extraction is the task of obtaining structured semantic relationships from unstructured text [16]. The extraction of biomedical information is usually done with unstructured scientific texts or electronic health records [17]. The main sub-tasks that are carried out are: entity recognition, relationship extraction and event identification [18].

In [19] one of the simplest methods to identify relationships between entities is applied. They used statistics of co-occurrences to calculate the degree of association between diseases and drugs in clinical records. However, most of the approaches based on co-occurrences achieve high recall and low accuracy.

The high degree of malformation in medical texts makes searches difficult, therefore it is convenient to integrate an information retrieval system into the relational databases [20].

In [21,22] we find a new method to process texts and represent them through intermediate forms that allow the related terms to remain together, thus preserving the semantics and allowing the identification of relationships between the terms. This form of textual representation is also based on co-occurrences, but is able to obtain high precision values.

In this study, we have gone into more depth in this method and applied it in the medical field to test its effectiveness and use in such a heterogeneous field.

1.1.2. Tag Cloud in Databases

In [23] a graphical interface for browsing and querying in a primary care database has been developed. The main advantage of this system is that all kinds of users would be able to handle it, including non-experts.

The tag cloud can have the same function but it wins in simplicity. The tags can be displayed within the tag cloud with different font sizes depending on their frequency of use [7]. These different font sizes give semantic expression to the tag cloud and highlight the most relevant terms [24].

Tagging is widely used in text mining in biomedical texts [18], in order to simulta-9neously obtain grammar and meaning. However, in databases, this type of collaborative tagging is not easily applicable. Some authors [25,26]have previously used tag clouds in databases to summarize query results. However, our main objective is to represent text content, facilitating the visual exploration of the information, although our tag cloud can also work as a query assistant.

Other authors [27,28] have represented the global content of the database through a tag cloud in which the tags are composed of single terms, whilst our tag cloud allows the use of multi-terms in tags.

2. Materials and Methods

In order to deal with the difficulties found in medical databases with unstructured texts described in Section 1, we propose the tag cloud as a visual interface. The tag cloud represents and summarizes the text, allows concept identification, their relevance and relationships, and facilitates browsing and querying.

The mathematical model proposed to define a multi-term tag cloud is based on the concept of the WAPO-Structure and its properties [29]. The WAPO-Structure preserves the semantics of the text by allowing related terms to remain ordered and together, a very important feature in the visualization of the tag cloud. The WAPO-Structure is generated from a method based on co-occurrence of sequences in the text [22]. Once generated, the structure represents the content of the information, serving as a tool for identifying concepts. Its visualization through the tag cloud allows non-expert users or those without prior knowledge of the text to make queries intuitively.

Next, we summarize the complete methodology followed to obtain a tag cloud from a textual attribute in a database.

2.1. Methodology

The methodology has been developed in the following steps: pre-processing, generation of the intermediate form, post-processing and visualization.

2.1.1. Pre-processing

First a syntactic preprocessing is carried out, applying tokenization filters, removing the stop words and performing a simple stemming to eliminate gender and plural forms from the raw data.

Subsequently, a semantic preprocessing is applied for which auxiliary files are created with the help of experts, containing the synonyms and acronyms of the specific hospital language.

For example, the term "IZQUIERDO", appears in the original text with all the following forms (considered synonymous in the auxiliary file): IZQ, IZDA, IZDO, IZQD, IZQDO, IZQDA, IZ, IZQIERDO, IZQU, IZQUI, IZQUERDO, HIZQ, IZQ1-UIERDA, IZQUIE, IZQUIEDO, IZQUIER, IZQUIERA, IZQUIERD, IZQUIERDOP, IZQUIERTDO, IZQUIRDO, IZQUUIERDO, IZQUIERDA. And this is without considering the different variants when mixed uppercase and lowercase occurs.

The acronyms file used fulfills the main function of obtaining all the acronyms referring to the same concept with the same syntactic form. Table 1 shows a sample of the acronyms in such file.

Table 1: Examples of acronyms found in the original text

Acronyms	Terms
EECC	Extracción extracapsular del cristalino
LIO	Lente intraocular
OI	Ojo izquierdo
EBA	Exploración bajo anestesia
CP	Cámara posterior
PA	Peritonitis aguda

The semantic pre-processing basically consists of removing the separation points between the letters in the acronyms and replacing the synonyms found in the original text by a canonical representative form (in the example seen for "IZQUIERDO", all those different forms of the term were replaced by a single one: "IZQUIERDO").

Table 2 shows an example of the clean text after applying the syntactic and semantic preprocessing.

Table 2: Example of the text after the cleaning

Short text	Modified short text
AMPUTACION 4 DEDO PIE IZDO	AMPUTACION 4 DEDO PIE IZQUIERDO
ENDORREDUCCION DE F.A.V.	ENDORREDUCCION DE FAV
legrado	LEGRADO
Mastectomia mas D.A.	MASTECTOMIA MAS DA

2.1.2. Generation of the Intermediate Form of Representation

In order to extract the WAPO-Structure [29], first the text is searched looking for frequent terms according to a given support. Once located, they combine with each other to give rise to the candidate sequences of two terms, those that are frequent in the text re-merge with each other, keeping the strict order of adjacency between terms, to give rise to the candidate sequences of three terms and so on until the frequent sequences with the greatest number of terms are found, which are called the spanning sequences.

WAPO-Structure.

Let $X = \{x_1, x_2, ..., x_n\}$ be a referential set of items and $S = \{A, B, ...\}$ a set of frequent weighted sequences with a cardinal higher than or equal to one, and $A, B, ...$ weighted AP-Sequences such as:

$$\forall\ A, B \in S; \quad A \nsubseteq B, \quad B \nsubseteq A \ \text{ and } \ B \neq A \quad . \tag{1}$$

A WAPO-Structure generated by S, $E = g(A, B, ...)$ is the set of AP-Sequences whose spanning sequences are $A, B, ...$

Note. If A is a spanning sequence of the WAPO-Structure, then all its sub-sequences will be frequent and belong to the WAPO-Structure.

Each AP-Sequence in the WAPO-Structure will be the generated from a spanning sequence as the following example shows.

Example of an AP-Sequence.

Let us suppose that the sequence {amputatión dedo pie izquierdo} is a spanning sequence, then a AP-Sequence is generated from it as shown in Figure1.

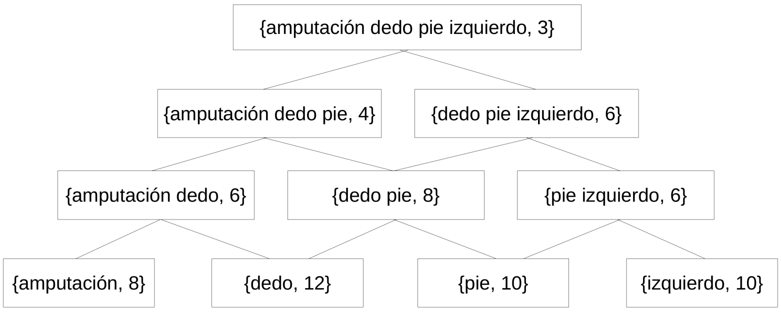


Figure 1. AP-Sequence generated from the spanning sequence {amputación dedo pie izquierdo}

The number that appears next to each sequence of terms indicates the absolute frequency (weight) of this sequence in the text.

The WAPO-Structure is the set of all the AP-Sequences generated from all the spanning sequences found in the text.

In order to decide the support value for our experiment, a group of experts belonging to our research group helped us to carry out a trial and error analysis. Small values for the support makes the number of terms in the visualization too large, which prevent them from being distinguishing clearly. For large values, important terms are lost in the visualization, so the most adequate support is that which keeps an acceptable number of terms in the visualization, but not so many as to make identification difficult.

Finally, we have selected the structure corresponding with a support equal to 0.3%.

WAPO-Structures provide some methods to manipulate them and their spanning sequences. The operations provided are described in detail in [29]. In addition, an alternative generation method is presented in [30].

2.1.3. Post-processing

Once all the information from the frequent itemsets is obtained, including their weight and the WAPO-Structure has been generated, we perform the post-processing. First, we use part of speech tagging. Nouns are good candidates to be represented in the tag cloud, but verbs and adverbs are not especially useful words to represent the content of information in this specific field. For this reason, in the post-processing stage, we remove all itemsets that contain verbs or adverbs. Adjectives by themselves are not informative, but if they are next to a noun, then they modify it adding semantics.

Considering these observations, a set of rules has been defined to determine when an itemset meets the requirements for its visualization within the tag cloud according to its grammatical category. The rules are the following for the Spanish language:

- *Level one itemsets*: The good candidates are nouns [N].
- *Level two itemsets*: The good candidates are those composed of two nouns [NN] or of an adjective and a noun: [AN] and [NA].
- *Level n itemsets*: The good candidates are those composed of a valid combination of terms in the previous levels plus a noun or an adjective. For level three it would be: [NNN], [ANN], [NNA], [NAN], [ANN], [AAN] and [ANA].

In addition to the frequent itemsets that contain verbs or adverbs, all those that do not meet any of these rules are removed. The rest are displayed in the corresponding tag cloud.

2.1.4. Visualization

After the post-processing, colour is applied to the tags in order to make it easier to distinguish them. We have chosen black for mono-terms and red shades for the multi-terms. In this way, an aesthetic tag cloud is achieved in which the colours have a functionality.

In Figure 2 we can see the visualization of the example in Figure 1.

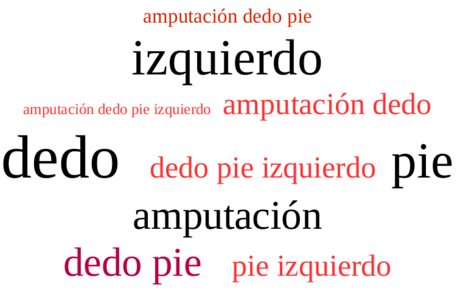


Figure 2. Visualization of the AP-Sequence in Figure 1

2.2. Dataset

For the experimental evaluation, we have used anonymous data from the Electronic Records of the "Hospital Clínico San Cecilio de Granada ", Spain. The data is in Spanish, and it is stored in a set of tables in a relational database. We have selected as a starting point the table that gathers together the information referring to the surgical interventions and that includes 24481 records. The main attributes of this table are " Diagnostics " and " Proposed Intervention ", where the content is a short text composed of one or more sentences.

We generate a tag cloud from the attribute "Proposed Intervention" (see Figure 3). This attribute is especially complex since it contains a large number of syntactic variations and the information in it was introduced by different practitioners in natural language.

2.3. Metrics for Evaluation

A retrieval system has been implemented where the tags work as queries. The textual records of the attribute have first been tagged with the most appropriate tags of the cloud in order to know which ones should be recovered with each of these tags. Then, the precision and recall can be calculated.

As time constraints do not allow us to manually tag the 24481 records, we set a relative error of 5% for a 95% of confidence in the mean estimation, which gives us a sample size close to 500 records that have been randomly selected.

The tagging has been carried out by a group of experts in the medical field. They have been provided with the records of the sample and a tag cloud generated from these records which contains the tags to use. It is important to highlight that the entire tag cloud generation process is automatic and this manual tagging is only performed to evaluate the tag cloud according to the opinion of the experts.

Next we calculate **coverage**, **overlap** and **balance** metrics with the formulas provided in [31]. The average coverage per tag has also been calculated, understood as the fraction of the original text represented, on average, by each tag in the tag cloud.

2.4. Satisfaction Survey

In order to evaluate the degree of user satisfaction with the tag cloud, we carried out a survey.

2.4.1. Procedure

We contacted several specialists requesting for collaboration by email. They were provided with a link to a website where we had enabled the tool for experimentation and subsequent evaluation. On this same website we embedded a form with some brief instructions and some statements with which the participants had to express their degree of agreement.

The first four statements were the following:

1. The tag cloud presented seems intuitive and easy to use.
2. The tag cloud presented provides information about the content of the database.
3. The information retrieved with the tags is consistent with them.
4. A tag cloud like the one presented would help me to search a medical database.

The last four are related to the ease to identify concepts. They are formulated in the following way:

"(5-8). It is easy for me to identify a concept in the tag cloud related to ...(*definitions in Table* 3)"

The reason for choosing these 4 concepts is because of their different locations in the visualization and their different sizes.

The first statement is related to the simplicity of the tag cloud, the second one to the content representation, the third and fourth to the retrieval system implemented over the tag cloud and the last four are for evaluating the ease of the identification of the terms.

doi:10.20944/preprints202107.0624.v1

Table 3: Concept definition and expected objects to be identified in the tag cloud in Figure 3

	Definitions provided in statements 5-8	Term
1	Surgical intervention for the birth of a baby	Cesárea
2	Surgical abortion or treatment after abortion	Legrado obstétrico
3	Surgical operation aimed at complete reconstruction of an obstructed or ankylosed joint	Artroplastia total
4	Technique widely used in the operation of cataracts	Facoemulsificación

The degree of agreement with these statements is expressed through a numerical rating from 1 to 5, where 1 indicates “complete disagreement” and 5 "complete agreement”.

Based on this classification, the correspondences of the degree of agreement with the given rate are:

- 1: Complete disagreement
- 2: Disagree
- 3: Indifferent
- 4: Agree
- 5: Complete agreement

Finally, the participants were asked to give some suggestions about the aspects to be improved.

2.4.2. Participants

The minimum sample size required to obtain an absolute error below 0.5 points in the mean estimation, considering a standard deviation equal to 1 and a 95% of confidence, is 18 participants. In total, we have information from 23 anonymous participants with training and experience in different areas of medicine, which has reduced the minimum error previously set.

3. Results

Table 4 shows the average of the precision, recall and $F1$ score considering all the tags in the generated cloud (see Figure 3). Two types of query have been considered for the calculation:

- **Type I**.- This retrieves all the entries where the terms appear in the same strict adjacency order they have on the tags.
- **Type II**.- This retrieves all the entries where the terms appear in the same order they have on the tags, without considering the strict adjacency.

Table 4: Average precision, recall and F_1 score for the tag cloud in Figure 3

Query types	precision	recall	F_1 score
Type I	0.940	0.780	0.823
Type II	0.990	0.870	0.904

The values obtained for precision, recall and F1 score are very good in both types of query, with type II being a little higher.

For precision, a higher value than for recall is achieved. This means that almost all the entries that are recovered are relevant, although there is a small proportion of relevant entries that are not recovered. We can solve this issue by increasing the support of the tag cloud, but an increase of the recall would probably decrease the precision. Furthermore, for recovering those records that remain inaccessible from the tag cloud, the system can also be queried in the traditional way.

Table 5 shows the values obtained for coverage, coverage per tag, overlap and balance.

To be such an extensive and heterogeneous dataset, a coverage close to 60% is a pretty good value. It is possible to increase this value by decreasing the support and

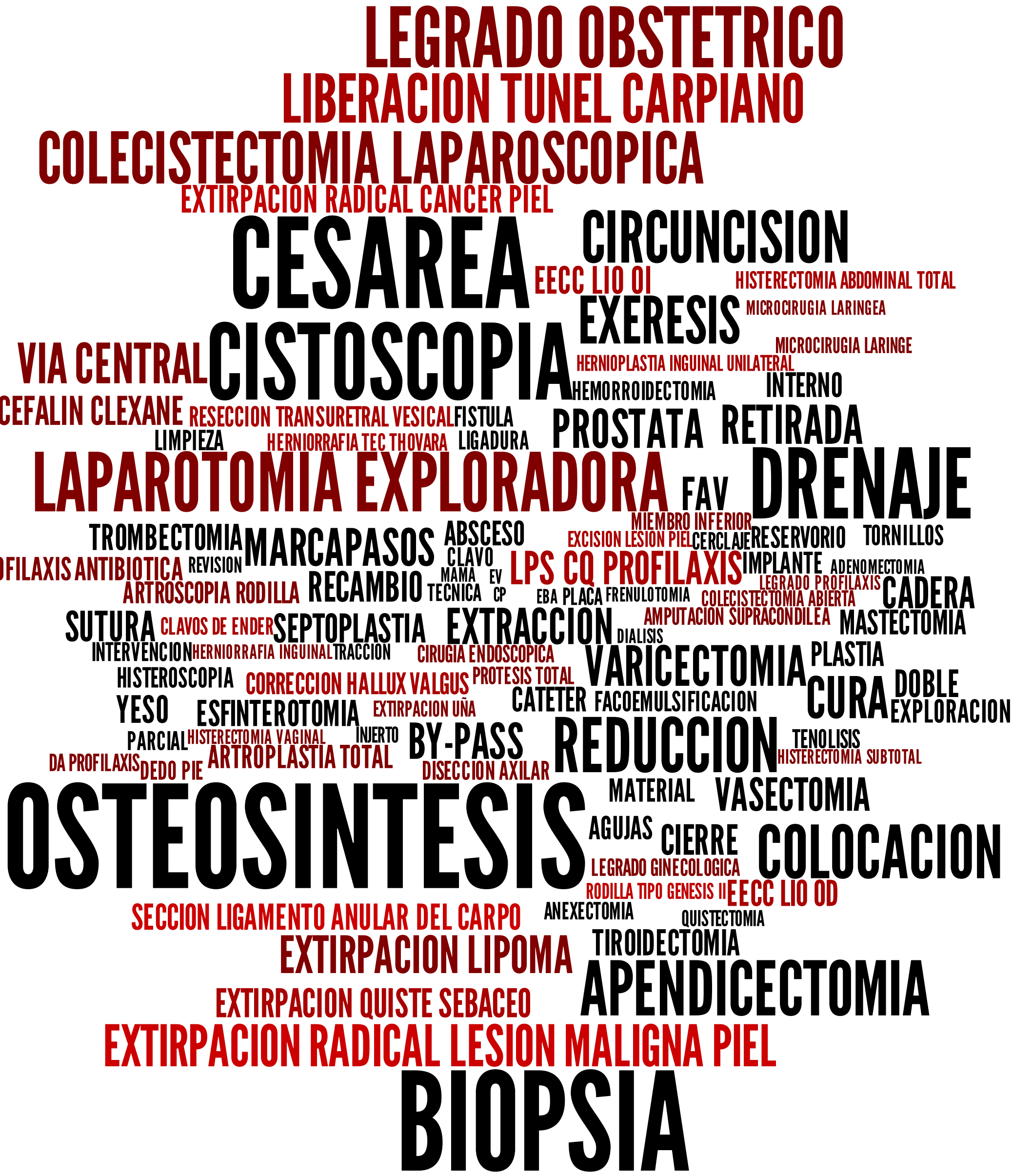


Figure 3. Tag cloud after postprocessing and colour with 0.3% of support

allowing a higher number of tags in the visualization. In other experiments we obtained better coverage with our automatic process, than other approaches where the tag cloud is manually built over the same database [29]. The overlap is practically 0, so different tags represent different information. Finally, the value obtained in the metric of balance indicates that the tag cloud is unbalanced, which gives this type of visualization the ability to highlight the most relevant topics.

3.1. Statistical Analysis of the Survey Results

The statistical analysis has been performed with StatGraphics Centurion XV.

The hypothesis that the assessments about the ease of concept identification are independent to the concepts provided is verified through a Chi-square independence test, with a discrepancy measure equal to 9.16 and a p-value equal to 0.6888. So the rates obtained for the four assessments regarding the concept identification are merged into only one variable.

Taking into account the rest of the assessments, there are in total five variables to analyze:

Table 5: Coverage, overlap and balance of the tag cloud in Figure 3

Coverage	Coverage per tag	Overlap	Balance
0.58793	0.00706	0.00003	0.05022

1. *Ease of use*: Evaluates if the tag cloud is intuitive and easy to use.
2. *Identification*: Evaluates whether the identification of concepts is simple in a global way.
3. *Retrieval*: Determines if the information retrieved is consistent with the tags.
4. *Representation*: Checks if the tag cloud represents the content properly.
5. *Utility search*: Evaluates if users would use the tool to search in medical databases.

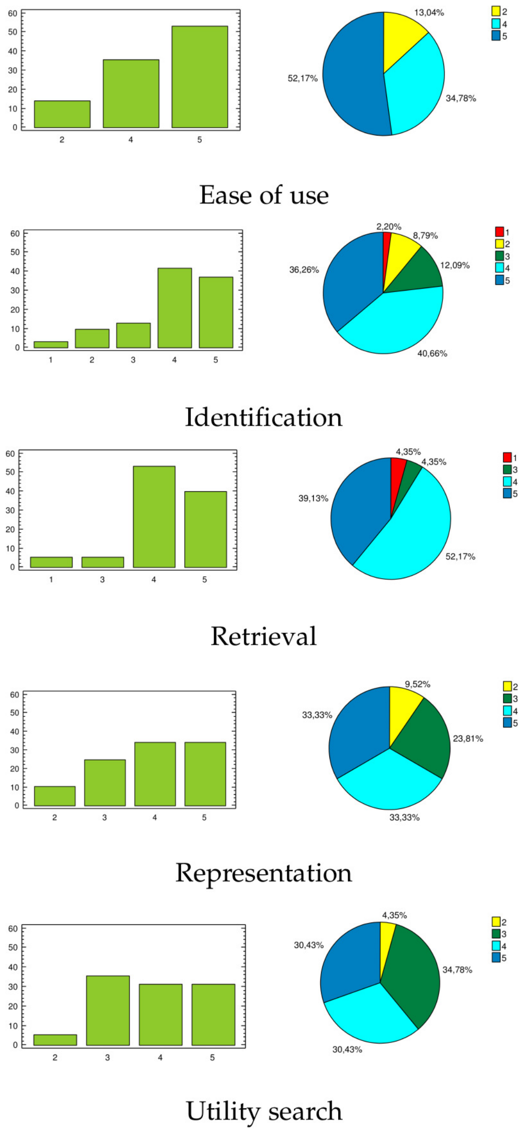
Figure 4 shows the bar graphs and the pie charts for each of these variables. When looking at the graphs, it is easy to realize that the highest percentages correspond to rates 4 and 5 (agree and complete agreement) for the first four variables. For the variable (agree and completely agree) for the first four variables. For the variable *Utility search*, the highest percentage is for rate 3 (indifferent), but this does not exceed the sum of the percentages corresponding to rates 4 and 5. In all cases, more than 60% of respondents give positive ratings (4 or 5) and this percentage is much higher for the first three variables, reaching almost 90% for *Ease of use* and surpassing it with *Retrieval*.

In the mean graph (Figure 5) the confidence intervals for the rates can be seen, with the one corresponding to *Ease of use* being the one with highest values. All the intervals have values greater than 3, which means that the average of the rates is greater than 3 in all cases for a 95% confidence, so the participants would agree with all the assessments made.

To verify this result, a unilateral test for the mean of a sample is performed for each of the five variables. In the null hypothesis we will have the mean being less or equal to 3 and in the alternative that is higher. In the case of rejecting the null hypothesis in favour of the alternative, we demonstrate that, on average, the participants agree with the statements made about the capacity of the tag cloud. To carry out this test, normality and equality of variances have been assumed. Table 6 shows the results summary.

Therefore, it can be said that the mean of all the variables is higher than 3, so the participants agree with the statements made about the capacity of the tag cloud with a significance of 5%:

1. Be intuitive and easy to use.
2. Represent the content appropriately.
3. Retrieve consistent information.
4. Help in carrying out searches.
5. The concept identification is easy.



Ease of use

Identification

Retrieval

Representation

Utility search

Figure 4. Bar graphs and the pie charts for each of the variables

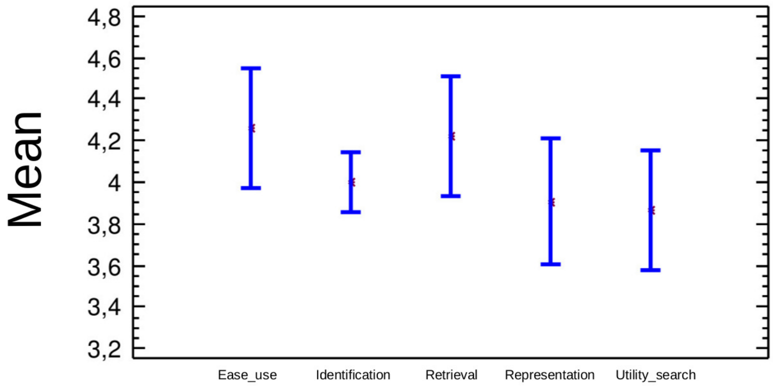


Figure 5. Mean graph of s with confidence intervals at 95% for the degree of agreement (from 1 to 5) with the different assertions about the characteristics of the tag cloud

Table 6: Summary result of the hypothesis tests

Verification that the mean of the degree of agreement with the assessments about the capabilities of the tag cloud can be considered higher than 3, that is, the participants agree with these assessments
Hypothesis test for the mean of a population: $H_0 : \mu \leq 3, H_1 : \mu > 3, \alpha = 5\%$ • *The tag cloud presented is intuitive and easy to use.* $t_{exp} = 5.98804$, p-value=$2.5 \cdot 10^{-6} \Rightarrow$ Reject $H_0 \Rightarrow$ Agree with the assessment. Sample mean = 4.26087 • *It is easy for me to identify the concepts provided in the tag cloud.* $t_{exp} = 9.33422$, p-value=0.0 $\Rightarrow$ Reject $H_0 \Rightarrow$ Agree with the assessment. Sample mean = 3.96739 • *The information retrieved with the tags is consistent with them* $t_{exp} = 6.47025$, p-value=$8.25 \cdot 10^{-7} \Rightarrow$ Reject $H_0 \Rightarrow$ Agree with the assessment. Sample mean = 4.21739 • *The tag cloud presented provides information about the content of the database* $t_{exp} = 4.16603$, p-value=$2.3 \cdot 10^{-4} \Rightarrow$ Reject $H_0 \Rightarrow$ Agree with the assessment. Sample mean = 3.90476 • *A tag cloud like the one presented would help me to search a medical databas* $t_{exp} = 4.5344$, p-value=$8.1 \cdot 10^{-5} \Rightarrow$ Reject $H_0 \Rightarrow$ Agree with the assessment. Sample mean = 3.86957

4. Discussion

We have applied a simple method based on co-occurrences to extract information from unstructured texts in a medical database and obtained very high precision which does not happen with other methods based on co-occurrences. This is due to taking into account and not only the isolated terms but also the sequences.

The $F1$ score obtained is similar to the best learning-based methods [17,32], both for the exact and the inexact matching, with our dataset being much bigger. We obtained more precision than these methods and less recall, but the recall values could be increased simply by adding a bigger number of terms to the tag cloud and decreasing the support. We have also improved the precision obtained in [21] by taking into account the order in the terms. The fact of not manually tagging the text in the tag cloud generation decreases the granularity found in other systems, such as the one seen in [33].

The coverage metric, close to 60% indicates that the tag cloud represents a good fraction of the original text, with the overlap being close to 0, which indicates that the tags represent different information. The balance is also close to 0 and is a good value for a tag cloud where the main objective is to be able to highlight the most relevant concepts, so the amount of results that are retrieved with each tag would be unbalanced. The coverage could be increased by increasing the number of terms in the tag cloud, but this would reduce their visibility and actually the value obtained is quite good considering the heterogeneity of the text. In addition, to recover those records that remain inaccessible from the tag cloud the system could be queried in the traditional way.

The satisfaction survey carried out with expert users verifies that, on average, they think that the tag cloud is an intuitive and user-friendly tool that provides information about the content of the database, which retrieves information consistent with the tags that compose it, which helps to search in a medical database and facilitates the identification concept. The great advantage of this tag cloud is that it does not require human intervention for its generation.

Limitations

We have only tested our approach on short texts, so it may not work properly on documents. Therefore, when the recall has to be increased, the support has to be increased, generating large tag clouds that may not be able to be displayed correctly on small screens.

5. Conclusions and Future Work

As we already know, it is difficult to process and access the information stored in large clinical databases and the graphical interfaces that help to query the database and to represent its content produces numerous deficiencies and are often complicated to use and the query results usually have low precision.

The main contributions of this study are to suggest a method for information processing that would work well in a large and heterogeneous clinical database, preserving the semantic of the textual attributes, and a graphical interface that is easy to use and properly represents the content information. This graphical interface helps to query the database and obtains query results with a high precision. We have provided a complete evaluation of the interface through the calculation of the appropriate metrics and a survey of expert users with satisfactory results.

Some of the limitations of our proposal are in relation with the screen resolution, the human visual system resolution as well as the limits of available computational resources. To deal with these limitations, we can apply some of the strategies proposed in [3] as future work. Furthermore, we will consider the generation of tag clouds in additional databases as well as create a multi-language system for the generation of tag clouds based on ontologies.

Another idea is to have a multi-level tag cloud using an ontology where choosing one item allows seeing others in the entire category, or narrowing down the query. This would reduce the size of the cloud and allow for a combination with something like a faceted search. It would also be interesting to offer a multi-lingual tool so that non-Spanish-speaking researchers could access the database. We are also contemplating the application of clustering techniques in our data as well as focusing on the search of entities through second-level text mining.

Author Contributions:

- Torres-Parejo: Conceptualization, formal analysis, investigation, validation
- Campaña: Data curation, methodology, software
- Vila: Project administration, Supervision
- Delgado: Project administration, Supervision

Funding: This work has been partially supported by:

- "FEDER project of the University of Granada" (Spain), B-TIC-145-UGR18.
- "I+D+I project of the Junta de Andalucía" (Spain), P18-RT-1765.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: We do not have permission to share the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yan, S.; He, L.; Seo, J.; Lin, M. Concurrent healthcare data processing and storage framework using deep-learning in distributed cloud computing environment. *IEEE Transactions on Industrial Informatics* **2020**, *17*, 2794–2801.
2. Algarni, A.; Ahmad, M.; Attaallah, A.; Agrawal, A.; Kumar, R.; Khan, R. A fuzzy multi-objective covering-based security quantification model for mitigating risk of web based medical image processing system. *International Journal of Advanced Computer Science and Applications* **2020**, *11*, 481–489.
3. Ketcheng, Q.; Wang, L. Research on Visual Data Mining Technology. Journal of Physics: Conference Series. IOP Publishing, 2021, Vol. 1748, pp. 1–7.
4. Liu, L.; Liu, L.; Fu, X.; Huang, Q.; Zhang, X.; Zhang, Y. A cloud-based framework for large-scale traditional Chinese medical record retrieval. *Journal of biomedical informatics* **2018**, *77*, 21–33.
5. Viégas, F.B.; Wattenberg, M. TIMELINES Tag clouds and the case for vernacular visualization. *Interactions* **2008**, *15*, 49–52.
6. Kuo, B.; Hentrich, T.; Good, B.; Wilkinson, M. Tag Clouds for Summarizing Web Search Results. Proceedings of the 16th International Conference on World Wide Web. ACM, 2007, pp. 1204–1205.
7. Deng, Y.; Denecke, K. Aspect-Oriented visualization of the health status: An example in treatment of cervical spine defect. MIE, 2016, pp. 18–22.
8. Agili, A.; Fabbri, M.; Panunzi, A.; Zini, M. Integration of a Multilingual Keyword Extractor in a Document Management System. Proceedings of the 6th International Conference on Language Resources and Evaluation, (LREC) 2008, 2008, pp. 1362–1366.
9. Don, A.; Zheleva, E.; Gregory, M.; Tarkan, S.; Auvil, L.; Clement, T.; Shneiderman, B.; Plaisant, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. Proceedings of the 16th ACM Conference on Information and Knowledge Management, (CIKM), 2007, pp. 213–222.
10. Watters, D. Meaningful Clouds: Towards a novel interface for document visualization. *Online Notes. University of Chicago.* **2008**.
11. Panunzi, A.; Marco, F.; Massimo, M. Integrating methods and LRs for automatic keyword extraction from open domain texts. Proceedings of the 5th International Language Resources and Evaluation, (LREC), 2006, pp. 1917–1920.
12. Zhou, D.; Zhong, D.; He, Y. Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine* **2014**, pp. 1–18.
13. Friedman, C.; Alderson, P.; Austin, J.; Cimino, J.; Johnson, S. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association* **1994**, *1*, 161–174.
14. Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S.; Wang, G. Data processing and text mining technologies on electronic medical records: a review. *Journal of Healthcare Engineering* **2018**, pp. 1–12.
15. Stewart, R.; Velupillai, S. Applied natural language processing in mental health big data. *Neuropsychopharmacology* **2021**, *46*, 252–253.
16. Zong, C.; Xia, R.; Zhang, J. Information extraction. In *Text Data Mining*; Springer, 2021; pp. 227–283.
17. Liu, F.; Chen, J.; Jagannatha, A.; Yu, H. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993* **2016**.
18. Simpson, M.; Demner-Fushman, D. Biomedical text mining: a survey of recent progress. In *Mining text data*; Springer, 2012; pp. 465–517.
19. Chen, E.; Hripcsak, G.; Xu, H.; Markatou, M.; Friedman, C. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association* **2008**, *15*, 87–98.
20. Fisk, J.; Mutalik, P.; Levin, F.; Erdos, J.; Taylor, C.; Nadkarni, P. Integrating query of relational and textual data in clinical databases: a case study. *Journal of the American Medical Informatics Association* **2003**, *10*, 21–38.
21. Martin-Bautista, M.; Martinez-Folgoso, S.; Vila, M. A new approach for representing and querying textual attributes in databases. *International Journal of Intelligent Systems* **2015**, *30*, 1021–1045.
22. Torres-Parejo, U.; Campaña, J.; Delgado, M.; Vila, M. MTCIR: A Multi-Term Tag Cloud Information Retrieval System. *Expert Systems with Applications* **2013**, *40*, 5448–5455.
23. Tate, A.; Beloff, N.; Al-Radwan, B.; Wickson, J.; Puri, S.; Williams, T.; Van Staa, T.; Bleach, A. Exploiting the potential of large databases of electronic health records for research using rapid

doi:10.20944/preprints202107.0624.v1

search algorithms and an intuitive query interface. *Journal of the American Medical Informatics Association* **2013**, *21*, 292–298.

24. Yang, L.; Li, J.; Lu, W.; Chen, Y.; Zhang, K.; Li, Y. The influence of font scale on semantic expression of word cloud. *Journal of Visualization* **2020**, *23*, 981–998.
25. Koutrika, G.; Zadeh, Z.; Garcia-Molina, H. Data Clouds: Summarizing keyword search results over structured data. Proceedings of the 12th ACM International Conference on Extending Database Technology: Advances in Database Technology, (EDBT), 2009, pp. 391–402.
26. Venetis, P.; Koutrika, G.; Garcia-Molina, H. On the selection of tags for tag clouds. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, (WSDM), 2011, pp. 835–844.
27. Deng, Y.; Denecke, K. Visualizing unstructured patient data for assessing diagnostic and therapeutic history. MIE, 2014, pp. 1158–1162.
28. Leone, S.; Geel, M.; Müller, C.; Norrie, M. Exploiting tag clouds for database browsing and querying. *Information Systems Evolution* **2011**, pp. 15–28.
29. Torres-Parejo, U.; Campaña, J.; Vila, M.; Delgado, M. A theoretical model for the automatic generation of tag clouds. *Knowledge and Information Systems* **2014**, *40*, 315–347.
30. Torres-Parejo, U.; Campaña, J.; Vila, M.; Delgado, M. Obtaining WAPO-Structure Through Inverted Indexes. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, 2018, Vol. 854, pp. 647–658.
31. Torres-Parejo, U.; Campaña, J.; Vila, M.; Delgado, M. Metrics for Tag Cloud Evaluation. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, 2018, Vol. 853, pp. 289–296.
32. Tang, B.; Cao, H.; Wu, Y.; Jiang, M.; Xu, H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC medical informatics and decision making* **2013**, *13*, 1–10.
33. Kang, T.; Zhang, S.; Tang, Y.; Hruby, G.; Rusanov, A.; Elhadad, N.; Weng, C. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association* **2017**, *24*, 1062–1071.