Chapter 2
A Recommender System of Medical Reports Leveraging Cognitive Computing and Frame Semantics

Danilo Dessì, Diego Reforgiato Recupero, Gianni Fenu and Sergio Consoli

Abstract During the last decades, a huge amount of data have been collected in clinical databases in the form of medical reports, laboratory results, treatment plans, etc., representing patients health status. Hence, digital information available for patient-oriented decision making has increased drastically but it is often not mined and analyzed in depth since: (i) medical documents are often unstructured and therefore difficult to analyze automatically, (ii) doctors traditionally rely on their experience to recognize an illness, give a diagnosis, and prescribe medications. However doctors experience can be limited by the cases they are treated so far and medication errors can occur frequently. In addition, it is generally hard and time-consuming inferring information for comparing unstructured data and evaluating similarities between heterogeneous resources. Technologies as Data Mining, Natural Language Processing, and Machine Learning can provide possibilities to explore and exploit potential knowledge from diagnosis history records and help doctors to prescribe medication correctly to decrease medication error effectively. In this paper, we design and implement a medical recommender system that is able to cluster a collection of medical reports on features detected by IBM Watson and Framester, two emerging tools from, respectively, Cognitive Computing and Frame Semantics, and then, giving a medical report from a specific patient as input, to recommend similar other medical reports from patients who had analogues symptoms. Experiments and results have proved the quality of the resulting clustering and recommendations, and the key role that these innovative services can play on the biomedical sector. The proposed system is

D. Dessì (✉) · D. Reforgiato Recupero · G. Fenu
Mathematics and Computer Science Department, University of Cagliari,
Via Ospedale 72, 09124 Cagliari, Italy
e-mail: danilo_dessi@unica.it

D. Reforgiato Recupero
e-mail: diego.reforgiato@unica.it

G. Fenu
e-mail: fenu@unica.it

S. Consoli
Philips Research, Data Science Department, High Tech Campus 34,
5656 AE Eindhoven, The Netherlands
e-mail: sergio.consoli@philips.com

G. A. Tsihrintzis et al. (eds.), *Machine Learning Paradigms*, Intelligent Systems Reference Library 149, https://doi.org/10.1007/978-3-319-94030-4_2

able to classify new medical cases thus supporting physicians to take more correct and reliable actions about specific diagnosis and cares.

Keywords Health recommender systems · Data mining · Cognitive computation Personal health records · Clustering · Knowledge inference · Personalized medicine · Relevance computation · Biomedical text-mining

2.1 Introduction

During the last decades a lot of data have been collected in textual clinical datasets representing patients' health states (e.g. medical reports, treatment plans, laboratory results, clinical records, surgical transcriptions, researches results etc.). Hence, digital data available for patient-oriented decision making has extremely grown but is not often mined and analyzed. Therefore, efficient access to information becomes hard for end-users [1]. In order to overcome text data overload and transform the text into useful and understandable source of knowledge, automated processing methods are required. Undoubtedly, this data can be exploited for figuring out relevant insights in the healthcare industry through data mining and machine learning techniques. These can work as a potential base for developing recommender systems which employ documents as items, and try to suggest diagnosis for new patients who present a clinical state similar to those that have been previously evaluated.

Recommender systems can be divided into three main categories: collaborative, content-based, and hybrid systems. Collaborative recommender systems work on experience gathered from previous user experiences, i.e. exploiting items which have been previously chosen by other users similar to a target in order to predict similar needs. Content-based recommender systems focus on the characteristics of items, e.g. when searching for a car, the recommendation output could be based on its price, brand, and color. Finally, hybrid recommender systems combine features of context-based and collaborative systems [2]. Hereafter, we focus on content-based recommender systems.

Content-based recommender systems usually rely on descriptions of people and items to build models which can be exploited for suggesting items similar to those a target person already had in the past [3]. They often employ retrieval approaches as a Vector Space Model (VSM), e.g. *bag-of-words*, as in [4]. A VSM is a model where each item is represented in a N-dimensional space and each dimension is related to a word of the documents collection. Many times, word-based approaches have not been able to figure out features for good results raising problems of accuracy. Therefore, data should be more deeply analyzed to yield a better understanding of users' state. One main challenge with medical reports is that a lot of information is stored by using the natural language, which suffers from the classical problem of ambiguity. Polysemy, troponymy, metonymy, n-grams expressions, entity recognition and disambiguation are common inherent problems of traditional methods largely employed in literature for dealing with textual resources. They make hard

to elaborate the contained information by means of machines, preventing the storing and sharing between different agents, processes and systems. As a consequence, recent studies have started to employ Semantic Web and knowledge based resources for obtaining better results.

New researches have introduced Semantic Web techniques combining ontologies and knowledge-based resources for shifting from word-based to concept-based representations of textual resources. This implies an increasing adoption of Semantic Web resources, tools and best practices for discovering the best features which play significant roles into unstructured texts, enabling high level categorization of contents. New systems, usually named Cognitive Computing systems, have earned a lot of attention for figuring out relevant insights from textual data. One system is IBM Watson[1] which can understand concepts, entities, sentiments, keywords, etc. from unstructured text through its Natural Language Understanding[2] service.

WordNet [5] and FrameNet [6], among others, are two of the most important linguistic open data resources that have been illustrated several times. WordNet is a lexical database that defines synsets as groups of synonyms. Each synset represents a unique meaning, which is semantically related to other meanings through derivation, hyponym/hypernymy, meronymy/holonymy, antonymy, entailment, etc. relations. FrameNet contains frames, which contextualize a general situation or state. Each frame includes semantic roles known as *frame elements* which are activated by lexical units of the speech (e.g. different verbs evoke different frames). However, its limited coverage and non-standard semantics are two major barriers for its wide adoption on natural language data analysis. To overcome these issues, a novel frame semantic tool, Framester [7], has been recently proposed. Framester works as a graph-linked data hub between open data systems as FrameNet, BabelNet [8] and WordNet, providing a dense interlinking between existing resources and enabling a novel formal semantics for frames. Framester can perform semantic frames and BabelNet synsets detection which may improve matchings between meanings of data expressed by different words. It is public available through an online interface[3] and an API.[4]

Technologies as Data Mining, Natural Language Processing, and Machine Learning can provide novel alternatives to explore and exploit potential retrieved knowledge from historical medical records, and help doctors to prescribe medication correctly to decrease medication errors effectively. In fact, text and data mining approaches have been already employed in healthcare for saving time, money and life [9–11].

Knowledge based techniques and tools, if reliable, can support medical staff in diagnosis, prevention and treatment of diseases, providing suggestions based on past medical cases. This chapter shows how to build a content-based recommender system within the healthcare domain leveraging Semantic Web technologies and cognitive computing tools.

---

[1] https://www.ibm.com/watson/.

[2] https://www.ibm.com/watson/services/natural-language-understanding/.

[3] https://lipn.univ-paris13.fr/framester/en/wfd_html/.

[4] https://github.com/framester/Framester/wiki/Framester-Documentation.

Moreover, we performed tests on a real dataset showing enhancements in embedding Semantic Web and Cognitive Computing tools. We examined which features better detect distinct characteristics from texts, and result suitable to cluster medical documents in order to provide high quality recommendations.

The chapter is organized as follows. First, we present the research on biomedical text analysis in Sect. 2.2. Then, we describe our recommender system in Sect. 2.3. In Sect. 2.4, we show our experiments and discuss the results we obtained. Finally, Sect. 2.5 proposes future development of our system and directions where we are headed.

2.2 State of the Art

It would be impossible to enumerate the numerous medical questions dealt with computational approaches for clinical enhancements. Here, we focus on an overview of the most interesting and promising text, data mining and machine learning methods, and their applications, to discover insightful information from textual data in order to support the development of a novel content-based recommender system.

2.2.1 Biomedical Information Retrieval

In recent years, many retrieval tools have appeared and have been used on textual resources for extracting relevant and insightful semantics [12, 13]. These tools usually exploit statistical techniques, even though there have been recently based on open linked data and machine learning techniques. Medical text processing is not a new question, but extracting biomedical data into a well-defined structural storage still remains a complex task [14]. Dealing with various medical domains does not help the development of systems to support medical activity. Because biomedical information is continuously being created in textual form more than ever before, there have been a lot of efforts for coding information into databases, and developing automatic processes which aim at finding useful ways to represent and organize data [15]. Medical text processing on medical domain, in particular using Natural Language Processing (NLP) approaches, has been explored into many other works [10, 11]. In general, researchers have usually tried to overcome text-depending issues focusing on classic entity recognition and text disambiguation techniques to create a domain-specific semantic content for the analysis of medical reports [14, 16, 17].

To alleviate textual inherit issues, some proposals have started to adopt Semantic Web practices in the medical system development. The first competition [18] on medical text-mining was run in 2002 during the Knowledge Discovery in Databases (KDD) Challenge Cup. Participants faced with a curation problem for assessing medical documents from the FlyBase dataset in order to determine whether a document should be curated based on the presence of experimental evidence of Drosophila

gene products. Exploiting Part-of-Speech (POS) tagging and semantic controls determined by examining the training documents and by focusing on figures captions, a collection of manually constructed rules obtained best results on the presence of experimental evidence for the document clustering [19]. In [20] the authors used a Support Vector Machine which was trained on MEDLINE abstracts to distinguish abstracts containing information on protein-protein interactions, prior to curate this information into their BIND database. They used a bag-of-words model with classification techniques and discovered that classifiers could minimize the number of abstracts that the practitioners employed to read by about two-thirds.

Authors in [21] have proposed a new concept-based model which exploits various text mining approaches and their combinations for improving text clustering. They propose a labeler which evaluates the semantic contribute of each word in sentences, outperforming traditional methods and discovering that the semantics is less sensitive to noise. More recent approaches are based on semantic analysis which enables learning more accurate features defined by means of external knowledge bases. In [22] authors make able systems to face with challenges by exploiting cultural and linguistic background knowledge for better interpreting unstructured documents and reasoning on their content. In [23] an item recommender system has been provided for recommendation tasks of various resources (e.g. movies and books) exploiting Word Sense Disambiguation techniques based on WordNet lexical ontology for mapping contents by means of synsets. Similar techniques are studied today in medical domain.

2.2.2 Biomedical Classification

In this section, we present classification methods which have been adopted for dealing with unstructured clinical notes over past years.

Classification is a fundamental component in the biomedical domain due to its widespread utility in applications such as medical diagnosis and identification of genetic causes of disease. In [20] authors exploit various classification techniques as described in Sect. 2.2.1. One more approach on MEDLINE documents was proposed by [24] where authors applied a semi-supervised spectral approach technique for clustering contents over two types of constraint: must-link constraints on document pairs with high (MeSH)-semantic or global-content similarities, and cannot-link constraints on those with low similarities. The authors proved the good performance of their new method on MEDLINE documents, improving performance of linear combination methods and several well-known semisupervised clustering methods.

Authors in [25] experiment multi-label classification techniques by means of combinations of bag-of-words models, and adopt time series and dimensionality reduction approaches on the MIMIC II dataset. In [26], authors implemented a Support Vector Machine classifier on n-gram features retrieved from clinical notes of the Beth Israel Deaconess Medical Center to identify the mechanical ventilation and diagnosis of neonatal and adult patients. A Convolutional Neural Network

classification approach has been proposed by [27] to build models which enable to generate context based representation of health related information at sentence level. Predefined disease labels have been adopted by [28] to classify free text clinical notes. They propose two techniques Sampled Classifier Chains (SCC) and Ensemble of Sampled Classifier Chains (ESCC), which extend their dataset with selected labels in order to obtain a relationship between disease and classification.

Performances of some classification methods applied on clinical notes have been recently evaluated in [29]. Authors focused on feature selection techniques investigating different approaches of transformation methods in order to improve the multi-label classification task. They report advantages of using filtering techniques and hybrid feature selection methods. One more recent work where classification methods have been evaluated is [30]. The best results have been obtained when a hierarchical approach to tag a document by identifying the relevant sentences for each label has been exploited.

2.2.3 Biomedical Clustering

In this section, we describe clustering methods applied to biomedical texts, and discuss recent works.

The clustering is the unsupervised task of finding groups of similar items by segmenting a collection into partitions called clusters, where items in the same cluster are more similar to each other than those in other clusters. In our work, biomedical text clustering items are medical reports. In general, document clustering can show various insights considering different levels of granularity of texts (i.e. clusters can be composed by whole documents, paragraphs, sentences or terms). In this case, the clustering can be employed as a tool for organizing and browsing documents in order to enhance the retrieval of information [31]. In biomedical domain, it could be essential to investigate patterns of a set of medical reports on features of different stuff so that similar patients can be treated concurrently in similar way.

An interesting medical document clustering has been proposed by [32] where authors exploited an ontology-based term similarity to index terms in a set of medical documents. They used a spherical k-means clustering algorithm on PubMed documents sets in order to evaluate the proposed similarity technique.

In [33] authors employed the KNN clustering method for evaluating a new similarity measure based on the semantic connection between words of a electronic medical report set. Authors in [34] performed cluster analysis on medical posts of online health communities for recognizing various types of content. They found that clusters can be associated to common categories as treatments, procedures, medications and so on. A framework based on clustering analysis has been developed by [35] for exploring health related topic automatically in online communities integrating data with medical domain specific knowledge.

Features as biomedical concepts and semantic relationships were identified with the help of ad-hoc ontologies for building a graph representation in order to enhance the recognition of categories by means of clustering techniques in [9].

2.2.4 Biomedical Recommendation

In literature, several systems refer to medicine for identifying active relations of new patients states with past ones, but few of them exploit natural language or text mining for accomplishing recommendation tasks. In [36], authors describe a recommendation procedure which uses similarity measures for finding relations between online users' health data and medical information of Wikipedia to increase patients' autonomy in their personal health. The task to predict future health risks by means of a recommendation technique has been proposed by [37], where authors developed an engine called CARE in order to predict the future diseases risks of patients. To provide more accurate and personalized doctor recommendations, authors in [38] mined emotions from previous users' ratings adopting a topic model technique for developing a system named iDoctor. To engender advances on health recommender systems, the ACM Conference on Recommender Systems hosted a workshop in years 2016 and 2017 where specific-purpose health recommender systems have been presented, but no one focused on textual resources as narrative medical reports. In addition, these systems deal with clinical data in order to provide specific online services which target patients as end-users, but there are not systems which exploit data for supporting diagnosis fruition and physicians' work.

2.2.5 Cognitive Computing and IBM Watson

With the terms Cognitive Computing systems we refer to those smart systems that learn at scale, can learn with purpose, and recently have modules for interacting directly with humans. They are being developed to reduce costs, increase efficiency, accelerate discovery, make essential connections in large amounts of data. With the rapid growth of the availability of massive amounts of data, Cognitive Computing systems provide new opportunities for augmenting human expertise in a broad range of domains. Embedding Cognitive Computing services in novel systems results fundamental for dealing with previous unmanageable issues. In medical domain, Cognitive Computing systems can play a relevant role for supporting activities of practitioners, providing understandable access to clinical data and enhancing the precision of the medicine. In our research, we have employed the most promising Cognitive Computing system, IBM Watson[5] which provides a cloud suite of services by means of the

[5] https://www.ibm.com/watson/.

IBM Cloud[6] platform for dealing with huge amount of data, and returns interesting features that can capture medical insights. More specifically, we employed the outcomes of the Natural Language Understanding service[7] which has been developed for analyzing textual data and, therefore, it is suitable for managing unstructured and narrative contents of medical reports.

2.2.6 Frame Semantics and Framester

Frame semantics is a linguistic theory that defines a meaning as a coherent structure of related concepts [39]. To relate various concepts, knowledge-based resources are usually employed as corner stones of semantic technological approaches. A content-based recommender system aware of semantics has the ability to interpret natural language texts and makes conclusions on their content. In order to embed frame semantics in our recommender system we exploits Framester, a novel data linked resource that works as a hub between linked open data systems as FrameNet, BabelNet and WordNet. It is a new frame-based ontological resource that leverages an inter-operable predicate space formalized according to frame semantics [6] and semiotics [40].

2.3 Architecture of Our System

In this section, we describe the modules of our system which needs proper techniques for representing items and comparing new and old users' states. An overview of our system is depicted in Fig. 2.1. The reader can see:

- *Medical Reports Collection*. This is the set of reports on which the system can learn about past clinical historical cases.
- *Content Analyzer Module*. The module takes as an input the collection of reports and the new report that the user (e.g. a physician) wants to evaluate. It embeds various resources for mining features from textual components of medical reports.
- *Represented Medical Reports*. This is the output given back by the Content Analyzer Module. The output is formatted so that machine learning algorithms can be easily applied.
- *Machine Learning Module*. This module implements a set of classification and clustering algorithms that are used for building models which describe patients' profiles.
- *Clinical Patients Profiles*. They are profiles that have been built by algorithms that had been employed in the Machine Learning Module.

[6] https://www.ibm.com/cloud/.

[7] https://www.ibm.com/watson/services/natural-language-understanding/.

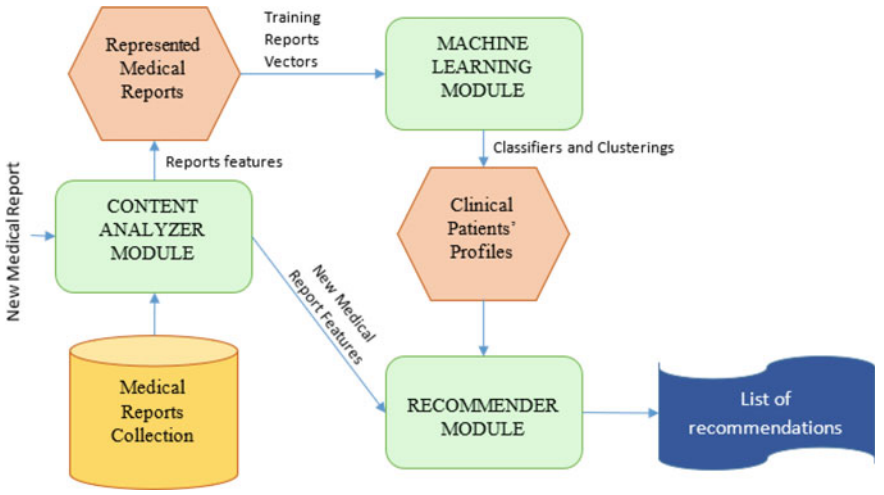

Fig. 2.1 Architecture of the content-based recommender system

- *Recommender Module*. The Recommender Module matches the new medical report features with the known patients' profile in order to make a list of recommendations.

2.3.1 Content Analyzer Module

The Content Analyzer Module takes as an input the collection of unstructured medical reports and produces a structured documents representation which enables the automatic computation of machine learning techniques performed by the Machine Learning Module. In addition, it must mine new unknown medical reports. In this section, the description of the model, the features and their characteristics are described.

2.3.1.1 Item Representation

For applying machine learning algorithms, data must be represented by sets of features usually called attributes. For example, to recommend books, attributes adopted to describe a book can be authors, editor, genre etc. When items are described by the same set of attributes and there are known values of these attributes, they are represented in structured data that can be employed for automatic computations. In case of biomedical textual documents there are not well-defined attributes, and textual features can raise difficulties when the system learns about patients. The main problem is that traditional term-based method can fail to capture the semantics of

clinical states of patients. For example, if more words can be used to indicate the same pathology (e.g. *tumors* could be indicated with the names *neoplasms, malignancies* etc.) relevant information can be lost if two clinical profiles do not contain the same word. In this context, semantic analysis of data plays a significant role and promises surprising results for solving these issues. More specifically, we have employed words coming from IBM Watson which is a leading Cognitive Computing tool, and Framester a novel hub between semantic resources. Subsequently, in this section we show features that can be extracted from medical textual resources and discuss about advantages of each one.

2.3.1.2 Vector Space Models

A Vector Space Model (VSM) is a spatial representation. For example, in a word-based VSM each document is represented over a N-dimensional space, where each dimension corresponds to a word that belongs to the whole set of terms of the given collection of documents. Let $D = \{d_1, d_2, \ldots, d_k\}$ be a collection of medical reports and $A = \{a_1, a_2, \ldots, a_n\}$ the set of attributes employed for representing them. A can be built by means of a natural language process or semantic content exploration pipeline which applies methods (e.g. the English stop word and stemming steps) for representing D. Each medical report d_i is represented by a vector of values $d_i = \{v_{1i}, v_{2i}, \ldots, v_{ni}\}$ where each value v_{ki} indicates the degree of relation between the attribute a_k and the document d_i. Attributes can have various natures such as words, n-grams, semantic features which describe contents, and so on. In our recommender system we have employed 6 different types of attributes: Term Frequency-Inverse Document Frequency, Concepts, Keywords, Entities, BabelNet Synsets, and Frames.

2.3.1.3 Term Frequency-Inverse Document Frequency

This is a *bag-of-word* model where attributes are words within the collection. For assigning a value to each word w, we have employed the Term Frequency-Inverse Document Frequency (TF-IDF) technique in which (i) uncommon words are not less relevant from frequent ones, (ii) a word that occurs many times in a document is not less relevant than a single one, and (iii) the length of documents does not play a significant role for the comparison of documents. To put it more simply, words that frequently occur within a document, but rarely in the whole collection, have more probability to be relevant in the document. The TF-IDF formula is showed in (2.1) where w_{ki} is the number of occurrences of the word w_k in the document d_i, $|d_i|$ is the size of the document expressed as number of words, N is the number of documents in the collection, and n_k is the number of documents where the word w_k occurs at least once.

$$TF - IDF(w_k, d_i) = \frac{w_{ki}}{|d_i|} \cdot log \frac{N}{n_k} \tag{2.1}$$

In order to prevent that longer texts have higher probability to be chosen by a recommender system, TF-IDF values are usually normalized in a range [0,1].

To avoid that frequent and no-relevant data (e.g. words that do not carry any meaning for the medical purpose as articles *the, a, an*, preposition *about, therefore, at*, etc.) appear in the TF-IDF features, the module performs some cleaning steps on the input texts. It precisely removes numeric data, punctuation, and stop-words. In fact, they are considered unnecessary and their remotion serves for (i) reducing the size of the VSM and (ii) for the subsequent efficiency of using a smaller space of features. All terms are taken in their lower case shape, avoiding to consider more times different representations of the same word (e.g. *Cardiac* and *cardiac*).

2.3.1.4 Concepts

Concepts can be defined as cognitive units which model perceived abstract subjects. They depend on the ability to process domain dependent knowledge and efficiently learn insights which become fundamental keys in the meaning of contents. Concepts can embody structures and representation of real words discovered in text, hence, they enable capturing high level abstraction reducing the complexity of the computation space. Moreover, they enable the specialization of employed attributes for representing documents in the VSM. IBM Watson can be employed for discovering automatically concepts related to the medical domain from natural language texts. It assigns a weight to each concept we have used for building the VSM. More precisely, given a collection of medical reports we use as set of attributes A the union of almost fifty concepts returned by IBM Watson from each medical report.

2.3.1.5 Keywords

Keywords are words of texts that enable listing the content of a report, releasing information about which words result relevant for describing the content of a document. Keywords are automatically detected by IBM Watson which provides a weight for each one. The VSM model is built as in the case of concepts.

2.3.1.6 Entities

Entities are actors that make actions in a text. Specifically to the medical domain, they can be people (e.g. physicians or nurses), illnesses (e.g. tumor), medicine names and so on. By capturing entities, it is possible to find relations between different documents if they share similar actors, especially when they are specific (for example if in a subset of documents D' physicians are cardiologists and in another subset D'' they are physiotherapists, the entities are distinct and enable better separation of the document subsets in different topics). As with the previous IBM Watson features, a weight is returned for each entity and indicates its influence in a document.

2.3.1.7 BabelNet Synsets

BabelNet synsets are unique unambiguous identifiers of sets of words which share the same meaning. We have chosen these synsets because (i) they are the result of the integration of various linguistic and semantic resources as WordNet, Wikipedia, FrameNet, among others, and (ii) they are directly provided by Framester. Differently from IBM Watson features, we do not have weights, hence, only the presence of BabelNet synsets have been considered by means of boolean flags into the Content Analyzer Module.

2.3.1.8 Semantic Frames

A semantic frame is a coherent group of concepts such that complete knowledge of one concept depends on the knowledge of all them in a context. Given a text, they are activated by nouns and verbs. Each frame can have multiple hierarchical levels that indicate its abstractions. For example, in Fig. 2.2, the word *cardiology* is abstracted by frames *Medical_specialties* and *Cure*. It should be underlined that frames are different from IBM Watson concepts because they do not depend on the application domain, but on relations that words have into linguistic and knowledge resources Framester adopts. As with the BabelNet synsets, we use the frames presence in the VSM.

2.3.1.9 The Course of Dimensionality Problem

The course of dimensionality problem refers to the issue that regards the great size of the number of attributes required to describe the target collection. The VSM suffers of this problem, hence, it needs to be managed into content-based applications as our

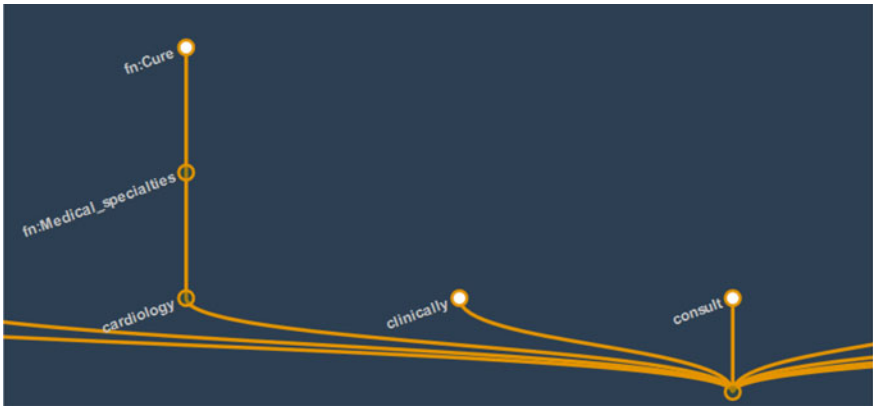

Fig. 2.2 Part of framester result on the sentence "*consider cardiology consult and further evaluation if clinically indicated*"

recommender system. One common method intensively applied in order to solve the issue is the Singular Value Decomposition (SVD). Let $A = \{a_1, a_2, \ldots, a_n\}$ be the set of attributes and $D = \{d_1, d_2, \ldots, d_i\}$ be the collection of our documents. The VSM is usually represented by a matrix M of size $|D| \times |A|$. M can be disjointed in three components $M = USV^T$ where S is a diagonal matrix containing the largest singular values, U is a matrix where columns are left singular vectors, and V is a matrix where columns define right singular values. In order to reduce complexity of data, the module applies a truncation which consists in holding only the largest k singular values, removing others which can be considered less relevant. This technique is known in literature as Truncated-SVD (TSVD). The module adopts the matrix $M' = U \times S$ which has a number of rows equivalent to the number of considered documents with a smaller number of attributes (columns) than the original matrix M. Besides decreasing the overall computational costs, an advantage of using the TSVD is deleting noise elements that might deteriorate the list of final recommendations. We want to point out that the value of k requires a trade-off between the amount of remaining and neglecting data to avoid the loss of information. Its value depends on the set of attributes A which characterizes the used collection.

2.3.2 Machine Learning Module

The Machine Learning Module receives a VSM as an input and returns a model which recognizes clinical patients' states. Its current version includes two clustering algorithms which are applied on all VSMs. The clustering techniques the module implements enable to deal with unsupervised data. They are (i) Hierarchical clustering algorithm and (ii) K-means clustering algorithm.

In this section we explain in depth how the chosen clustering algorithms work, and discuss about advantages they enable. Moreover, we show which machine learning algorithms might be employed in our system underlying which are requirements for an enhanced recommendation.

2.3.2.1 Hierarchical Clustering

Hierarchical Clustering builds a clusters hierarchy, or in other words, a tree of clusters which is usually called *dendrogram*. Each cluster contains children that are clusters as well, unless for the leafs of the tree. Sibling clusters split documents that are contained in the common parent cluster. A hierarchical clustering algorithm can be either *agglomerative* or *divisive*. In its agglomerative version, the algorithm starts with single elements of the collection, then it merges elements together based on a chosen measure (e.g. *Euclidean distance*). The agglomerative process is iterated as long as a unique cluster that covers all collection documents is obtained. The divisive variant of the algorithm starts with one cluster of all documents and recursively splits the most appropriate clusters according to a given criteria (e.g. splitting the largest

cluster in each iteration.). The method continues its execution until a stop criterion (e.g a given number of clusters) is achieved. Our recommender system implements an agglomerative clustering, since we are interested in building groups looking for similarities starting from pairs of documents.

The hierarchical clustering algorithms are easily applicable on each kind of data, enable a manageable granularity of clusters and can be applied with any type of similarity measures. For these reasons, we felt that this type of clustering approach can lead good results on medical domain.

2.3.2.2 K-means Clustering

The K-means Clustering is a partition method. It builds a set of clusters minimizing the sum of squared distance between elements of a cluster and its center. The results is a single partition of data without any structure and, hence, can have advantages on applications which involve large sets of data for which the construction of a hierarchical structure can be onerous. The algorithm requires the number of clusters k as an input. This number is used to allocate k random centers which will be employed to build clusters. At beginning, it assigns each element to the cluster with the nearest center. Iteratively, centers are updated based on the built clusters and elements are moved into the cluster with their nearest center.

2.3.2.3 Similarity Measures

Precise clustering requires an accurate definition of the closeness between documents represented in the VSM. The closeness can be measured by either the pair-wised similarity or distance. A variety of similarity or distance measures have been proposed and discussed in literature. Our Machine Learning Module adopts the Cosine and Euclidean measures. The Cosine similarity quantifies the angle between two documents expressed by vectors. Its formula applied on two vectors v_p and v_q can be observed in (2.2).

$$CosS(v_p, v_q) = \frac{v_p\, v_q}{\| v_p \| \| v_q \|} \tag{2.2}$$

$CosS$ values 1 when v_p and v_q are completely similar, and 0 otherwise.

The Euclidean distance $EucD$ between two vectors v_p and v_q is defined as usual in (2.3).

$$EucD(v_p, v_q) = \sqrt{\sum_i (v_p(i) - v_q(i))^2} \tag{2.3}$$

Differently from the Cosine similarity, the Euclidean distance has not a limited range of values, therefore, it needs to be scaled before used for the similarity

evaluation. For such reason the module adopts the formula (2.4) for scaling Euclidean-based values.

$$EucS(v_p, v_q) = \frac{1}{1 + EucD(v_p, v_q)} \tag{2.4}$$

2.3.3 Recommendation Module

This module uses the clinical patients' profiles for suggesting possible past medical cases that are similar to the new one by matching the new medical case against clinical profiles' clusterings of medical reports to be recommended. More specifically, the Recommendation Module takes a new medical report representation r and predicts whether there are clinical patients' profiles $p_1, \ldots, p_n$ that are interesting according to the relevance with r. It performs strategies to rank documents, and top-ranked ones are included in the list of recommendations that are provided to the final user. For doing so, the module computes the closeness between a new medical reports and clusters. In detail, given a new patients' medical report r and a clustering $C = \{c_1, \ldots, c_n\}$, the module finds the cluster c_i which has the closest center to r. Then elements within c_i are ranked from the most to the least similar to r. The produced ranking is used for finding the closest k medical reports as the final recommendation list.

2.4 Experiments

2.4.1 The Test Dataset

The employed dataset is a collection of no-labeled medical reports. It is freely available from the open-source iDASH repository.[8] In the dataset there are 2362 reports written in English. On the average, each report contains 400 words (ranging from 138 words for the shortest document to 1048 words for the longest one). There are singleton medical reports which might not have similarities with others, hence, for avoiding making unclear clustering groups they should be placed in one-element clusters.

Reports can be medical transcription samples including clinical notes, care plans, medical examinations, radiology reports etc. In the dataset, categories, their amount and distribution across reports, are not explicitly reported, although categories can be deduced from reports content (e.g. there are reports concerning heart issues whose can be placed in a category *heart*). The lack of predefined schema and the wide vocabulary of used terms make hard the categorization. Moreover, file names refer

[8] https://idash-data.ucsd.edu/.

to specific diseases or body parts issues that could be exploited to classify directly contents, but there could be ambiguous terms which may or may not refer to the same disease. Examples of medical reports names which can involve the same topic and discuss about the same issue are *cardiac-catheterization* and *hearth-catheterization*. As a consequence, these medical reports might be inappropriate to test supervised approaches, but they are suitable to test our unsupervised system which can manage unlabeled data.

2.4.2 Experiment Setup

2.4.2.1 Data Cleaning

Data cleaning is necessary in order to provide the same valid English text to the Content Analyzer Module services. First of all, we have cleaned all medical reports from HTML tags, removed all tables and structured format styles in order to obtain simple plain texts. Then we have matched reports words against those provided by WordNet, sending the word w' and getting the word w'' which has been placed in the text. At the end, only English text with correct grammar and punctuation composes the collection of medical reports.

2.4.2.2 Content Analyzer Module Setup

The Content Analyzer Module has been configured for providing more VSMs models which have been built on various features as described in Sect. 2.3.1. More precisely, let r_i be the i-th medical report and f_j be the j-th feature of a selected type. The outcomes of the module are:

- **5 Binary VSMs**: they include a matrix representation for the Concepts, Keywords, Entities, BabelNet Synsets and Semantic Frames features. Binary means that if f_j occurs within the inferred set of features of the medical reports r_i, in the VSM model M their relation is indicated by $M[i, j] = 1$, otherwise $M[i, j] = 0$;
- **4 Weighted VSMs**: they include a matrix representation for the Concepts, Keywords, Entities, and TD-IDF features. Weighted means that $M[i, j] = weight$, where $weight$ has been calculated exploiting the Natural Language Understanding service of IBM Watson or the TF-IDF approach as described above, and represents how strong is the relation between the medical report r_i and the feature f_j, otherwise $M[i, j] = 0$;
- **5 Counted VSMs**: they include a matrix representation for the Concepts, Keywords, Entities, BabelNet Synsets and Semantic Frames features. Counted means that $M[i, j] = count$ where $count$ is the number of times that a feature f_j occurs within the set of features of the medical reports r_i, otherwise $M[i, j] = 0$;

(a)

Report Name	Myocardial infarction	Heart	Atherosclerosis	Obesity	Cardiology	Cardiovascular system	Atheroma	Hypertension
heart-catheterization-ventriculography-angiography	1	1	1	0	1	1	0	0
cardiac-catheterization	1	1	1	0	1	0	1	0
cardiovascular-letter	1	0	1	1	0	0	0	1

(b)

Report Name	Myocardial infarction	Heart	Atherosclerosis	Obesity	Cardiology	Cardiovascular system	Atheroma	Hypertension
heart-catheterization-ventriculography-angiography	0.97	0.95	0.87	0	0.68	0.60	0	0
cardiac-catheterization	0.96	0.62	0.51	0	0.50	0	0.53	0
cardiovascular-letter	0.96	0	0.85	0.25	0	0	0	0.94

(c)

Report Name	Myocardial infarction	Heart	Atherosclerosis	Obesity	Cardiology	Cardiovascular system	Atheroma	Hypertension
heart-catheterization-ventriculography-angiography	4	6	4	0	4	2	0	0
cardiac-catheterization	3	3	2	0	2	2	2	0
cardiovascular-letter	4	0	3	3	0	0	0	3

Fig. 2.3 Samples of VSMs built on concepts extracted from three medical reports. Samples are related to **a** binary **b** weighted and **c** counted VSM

For more details on the three mentioned distances, the reader can look at examples of VSMs built on concepts extracted from three medical reports of the test dataset in Fig. 2.3. In the first row of each VSM, there are concepts that form the N-dimensional space. In the other rows, there are the names of reports on the first columns followed by values that indicate the degree of relation between the medical report and the i-th concept. The reader notices that (a) is built using the binary relation, (b) is built using the weighted relation and (c) is built using the counted relation.

2.4.2.3 Machine Learning Module Setup

The Machine Learning Module applied both clustering methods on all VSMs. In order to obtain high quality clusters, we set the module for exploiting the Silhouette width measure. Given a cluster c, its Silhouette width value $s(c)$ is computed as showed in Eq. (2.5) where $w(c)$ is the average dissimilarity within c and $o(c)$ is the lowest average dissimilarity of c to any other cluster.

$$s(c) = \frac{o(c) - w(c)}{max\{o(c), w(c)\}} \tag{2.5}$$

Values of Silhouette width range from -1 to 1. When the value is closer to 1, it means that the clusters are well separated; when the value is closer to 0, it might be difficult to detect the decision boundary; when the value is closer to -1, it means that

elements assigned to a cluster might have been erroneously assigned. In general, we can consider good clusterings those that have high average values of Silhouette width. Unsurprisingly, the value of the Silhouette width depends on the type of features of the VSM under processing.

Hierarchical clustering. After the hierarchical clustering has been computed, the resulting dendrogram has been iteratively cut starting from its head, in order to increase the number of clusters for each iteration. In doing so, various clusterings obtained with different cut values have been produced. As a reminder, in our dataset we do not know how many groups can be formed. Therefore, we have exploited the highest value of average Silhouette width values in order to cut dendrogram where the clustering showed the best separation between medical reports.

K-means clustering. K-means Clustering has been performed with different values of k as number of clusters. For each value of k, the average Silhouette width measure has been computed similarly to hierarchical clustering setup. Then the clustering with the highest average value of Silhouette width has been hold as the output of the module.

2.4.3 Recommendation Module Setup

The recommendation module has been setup to receive an unknown medical report and a number k which represents the number of recommendations. In our experiments the adopted value of k is 10.

2.4.4 Results

At the current state, the quality of results of our recommender system mainly depends on the Content Analyzer Module and Machine Learning Module. In fact, a good quality of clusters means that medical reports similar to a new one can be correctly detected in the test dataset. First, for obtaining good clustering the features must allow a good separation of reports, and second the clustering algorithm must recognize which the best divisions are. Therefore, in this section we discuss about the most representative features of our dataset and the clustering algorithms performance. Results of clusterings quality can be observed in Figs. 2.4, 2.5, 2.6, 2.7 and 2.8.

The sets of features which have formed the best division of medical reports into clusters are those that have been computed using IBM Watson. In fact, they reach good levels of silhouette width. In more details, concepts and entities in their weighted and binary mappings have showed good performances in capturing medical information from medical reports of the test dataset. This fact suggests that the relevance of an entity or a concept into a medical report does not depend on the number of times that it appears. We can say that their role depends on the relations they have into reports, and

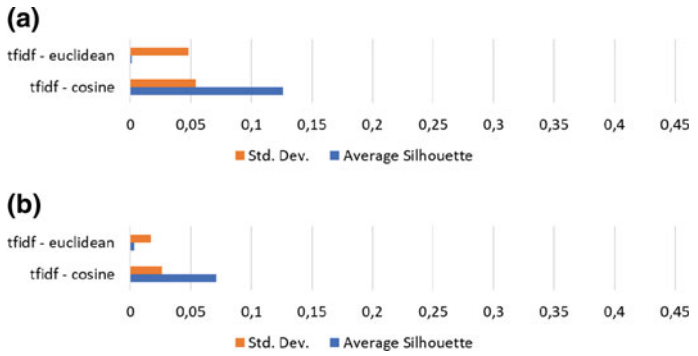

Fig. 2.4 The average and standard deviation values of the silhouette width measure of the clusterings computed on the TF-IDF measure. **a** Hierarchical clustering. **b** K-means clustering

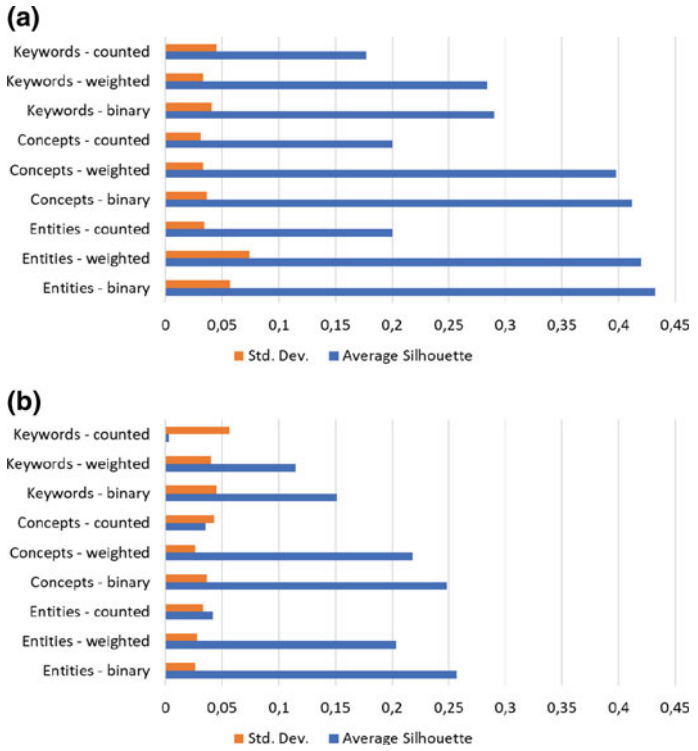

Fig. 2.5 The average and standard deviation values of the silhouette width measure of the clusterings computed on IBM Watson features. **a** Hierarchical clustering on cosine distance. **b** Hierarchical clustering on Euclidean distance

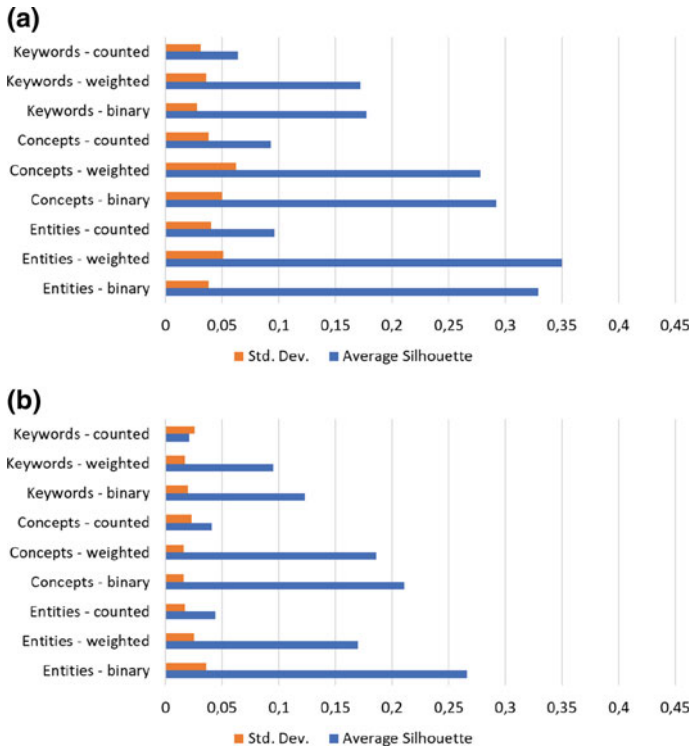

Fig. 2.6 The average and standard deviation values of the silhouette width measure of the clusterings computed on IBM Watson features. **a** K-means clustering on cosine distance. **b** K-means clustering on Euclidean distance

more influent their actions are, stronger their relevance is. Considering the number of times that a concept or an entity appears we do not add any additional information in our representative VSM. Keywords do not have showed good performances like entities and concepts but they could be considered as good alternatives in those cases where detecting entities and concepts can be hard.

Framester features do not have reached good results in the clusterings. This can depend on the fact that they are more abstract and not directly connected to the medical domain. Moreover, our test dataset could negatively influence this types of features since medical reports are strongly specific on patients' medical states. By contrast, they can result useful for those medical reports that describe the state of patients more in general without too clinical details (e.g. a starting examination visit). As for Framester features, the TF-IDF does not have showed good performances and same motivations can be observed.

One more point to consider is how the distance between two medical reports is computed. Results suggest that the cosine distance is more reliable than the Euclidean distance. Nevertheless, it is important underlying how they seem keeping a similar

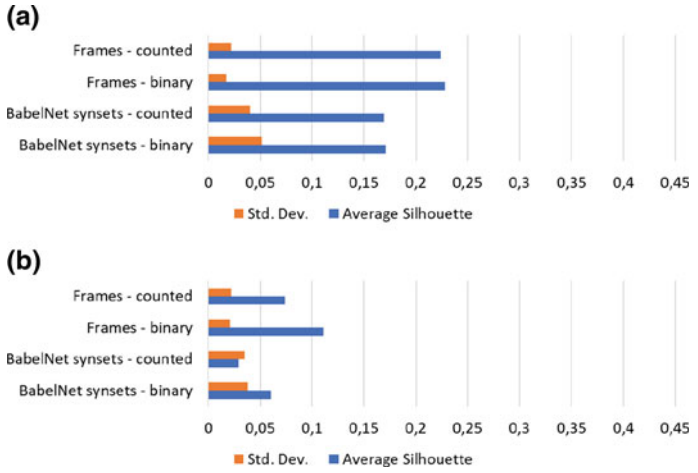

Fig. 2.7 The average and standard deviation values of the silhouette width measure of the clusterings computed on Framester features. **a** Hierarchical clustering on cosine distance. **b** Hierarchical clustering on Euclidean distance

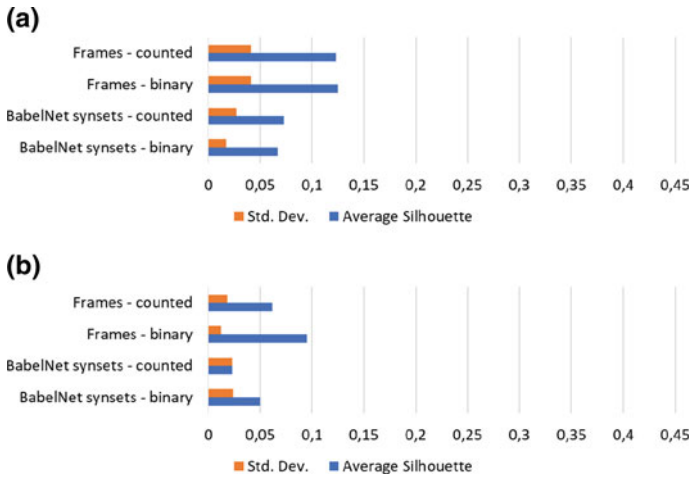

Fig. 2.8 The average and standard deviation values of the silhouette width measure of the clusterings computed on Framester features. **a** K-means clustering on cosine distance. **b** K-means clustering on Euclidean distance

Fig. 2.9 An example of list of recommendations built by our recommender system using as new report that called *heart-catheterization-angiography-1*

```
heart-catheterization-angiography-2
coronary-ct-angiography-ccta-2
heart-catheterization-ventriculography-angiography-6
stenting
cardiac-catheterization-8
heart-cath-coronary-angiography
nuclear-cardiac-stress-report
gen-med-consult-4
heart-catheterization-ventriculography-angiography-4
cardiac-catheterization-1
```

behavior on different features. To name an example, *entities-binary* and *entities-weighted* show a similar behavior both for cosine and for Euclidean distance.

Hierarchical clustering have outperformed the results of the K-means clustering, hence, if the recommender system would have been employed on a real medical case, the hierarchical clustering should be used. The agglomerative approach seems to be more suitable for finding medical cases similar to a new one.

Finally, to show how recommendation module has worked the reader can look at example in Fig. 2.9. The figure lists 10 medical reports of our test dataset that are returned by our recommender system when the report called *heart-catheterization-angiography-1* has been adopted for the evaluation of a new clinical state of a patient. The example shows how returned recommendations are correlated to heart issues and, hence, that our approach in building a recommender system can effectively recognize medical contents in order to suggest relevant past clinical cases.

2.5 Conclusion and Future Trends

Recommender systems are employed in many fields to help users to find important products and services for them. Similar approaches can be headed for providing diagnosis, thus supporting physicians in their work. In this chapter we presented a content-based recommender system within the medical domain, by providing an overview of recent information retrieval and semantic enrichment tools we employed. Our work addressed the challenge to find out which types of information can be directly processed by machines on large collections of medical reports, combining emergent cognitive computing systems in order to return reliable recommendation results. We discussed about the quality of features related to the representation of the medical reports content, underlying how they can capture the semantics from unstructured texts.

Subsequently, we discuss about two clustering approaches our recommender system currently implements. We used them to handle with various VSMs and explained their advantages and uses based on the type of dataset. In order to deal with classes of reports, the Machine Learning Module can be integrated with classification approaches and we aim at finishing this improvement in the immediate future. At the moment, we have not considered this evolution since it results hard to find datasets for a significant validation of classification tasks. We would like to underlie how recommender systems are substantial opportunities to progress in data science for the health-care field. For doing so, new resources and open datasets are required to enable further improvement of methods and designation of algorithms in clinical context.

Acknowledgements Danilo Dessì gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014–2020—Axis III Education and training, Thematic goal 10, Priority of investment 10ii, Specific goal 10.5).

References

1. Mishra, R., Bian, J., Fiszman, M., Weir, C.R., Jonnalagadda, S., Mostafa, J., Del Fiol, G.: Text summarization in the biomedical domain: a systematic review of recent research. J. biomed. Inform. **52**, 457–467 (2014)
2. Sezgin, E., Ozkan, S.: A systematic literature review on health recommender systems. In: IEEE E-Health and Bioengineering Conference (EHB), pp. 1–4 (2013)
3. de Gemmis, M., Lops, P., Musto, C., Narducci, F., Semeraro, G.: Semantics-aware content-based recommender systems. In: Recommender Systems Handbook, pp. 119–159. Springer (2015)
4. Capelle, M., Hogenboom, F., Hogenboom, A., Frasincar, F.: Semantic news recommendation using wordnet and bing similarities. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 296–302. ACM (2013)
5. Lin, D.: Review of "WordNet: an electronic lexical database" by Christiane Fellbaum. The MIT Press 1998. Comput. Linguist. **25**(2), 292–296 (1999)
6. Baker, F.C., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, ACL '98 and 17th International Conference on Computational Linguistics, vol. 1, pp. 86–90. Association for Computational Linguistics, Stroudsburg, PA, USA (1998)
7. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: a wide coverage linguistic linked data hub. In: 2016 20th International Conference on Proceedings of Knowledge Engineering and Knowledge Management, EKAW, pp. 239–254. Springer (2016)
8. Navigli, R., Ponzetto, S.P.: Babelnet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
9. Bleik, S., Mishra, M., Huan, J., Song, M.: Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(5), 1211–1217 (2013)
10. Cohen, A.M., Hersh, W.R.: A survey of current work in biomedical text mining. Brief. bioinform. **6**(1), 57–71 (2005)
11. Toor, R., Chana, I.: Application of IT in healthcare: a systematic review. ACM SIGBioinform. Rec. **6**(2), 1–8 (2016)
12. Presutti, V., Consoli, S., Nuzzolese, A.G., Recupero, D.R., Gangemi, A., Bannour, I., Zargayouna, H.: Uncovering the semantics of wikipedia pagelinks. In: Lecture Notes in Computer Science, vol. 8876, pp. 413–428 (2014)
13. Presutti, V., Nuzzolese, A.G., Consoli, S., Gangemi, A., Recupero, D.R.: From hyperlinks to semantic web properties using open knowledge extraction. Semant. Web **7**(4), 351–378 (2016)
14. Lushnov, M., Safin, T., Lapaev, M., Zhukova, N.: Medical text processing for SMDA project. In: EMSA-RMed@ESWC (2016)
15. Consoli, S., Stilianakis, N.I.: A quartet method based on variable neighbourhood search for biomedical literature extraction and clustering. Int. Trans. Oper. Res. **24**(3), 537–558 (2017)
16. Chernyshevich, M., Stankevitch, V.: IHS-RD-BELARUS: clinical named entities identification in French medical texts. Physiology **279**, 291 (2015)
17. Dessì, D., Recupero, D.R., Fenu, G., Consoli, S.: Exploiting cognitive computing and frame semantic features for biomedical document clustering. In: Proceedings of the Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics co-located with 14th Extended Semantic Web Conference, SeWeBMeDA@ESWC 2017, pp. 20–34 (2017)
18. Yeh, A.S., Hirschman, L., Morgan, A.A.: Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. Bioinformatics **19**(Suppl. 1), 331–339 (2003)
19. Regev, Y., Finkelstein-Landau, M., Feldman, R.: Rule-based extraction of experimental evidence in the biomedical domain: the KDD cup 2002 (task 1). ACM SIGKDD Explor. Newslett. **4**(2), 90–92 (2002)
20. Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., Hogue, C.W.V.: PreBIND and textomy—mining

the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinform. **4**(1), 11 (2003)
21. Shehata, S., Karray, F., Kamel, M.: An efficient concept-based mining model for enhancing text clustering. IEEE Trans. Knowl. Data Eng. **22**(10), 1360–1371 (2010)
22. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Recommender Systems Handbook, pp. 73–105. Springer (2011)
23. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. User Model. User-Adapt. Interact. **17**(3), 217–255 (2007)
24. Gu, J., Feng, W., Zeng, J., Mamitsuka, H., Zhu, S.: Efficient semisupervised MEDLINE document clustering with MeSH-semantic and global-content constraints. IEEE Trans. Cybern. **43**(4), 1265–1276 (2013)
25. Bromuri, S., Zufferey, D., Hennebert, J., Schumacher, M.: Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms. J. Biomed. Inform. **51**, 165–175 (2014)
26. Marafino, B.J., Davies, J.M., Bardach, N.S., Dean, M.L., Dudley, R.A., Boscardin, J.: N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J. Am. Med. Inform. Assoc. **21**(5), 871–875 (2014)
27. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. Stud. Health Technol. Inform. **235**, 246–50 (2017)
28. Zhao, R.W., Li, G.Z., Liu, J.M., Wang, X.: Clinical multi-label free text classification by exploiting disease label relation. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 311–315. IEEE (2013)
29. Glinka, K., Woźniak, R., Zakrzewska, D.: Improving multi-label medical text classification by feature selection. In: 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 176–181. IEEE (2017)
30. Baumel, T., Nassour-Kassis, J., Elhadad, M., Elhadad, N.: Multi-label classification of patient notes a case study on icd code assignment. CoRR abs/1709.09587 (2017)
31. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A Brief Survey of Text Mining: classification, clustering and extraction techniques. arXiv:1707.02919 (2017)
32. Zhang, X., Jing, L., Hu, X., Ng, M., Xia, J., Zhou, X.: Medical Document Clustering using Ontology-based Term Similarity Measures (2008)
33. Zhang, Y., He, Z., Yang, J.J., Wang, Q., Li, J.: Re-structuring and specific similarity computation of electronic medical records. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 230–235. IEEE (2017)
34. Chen, A.T.: Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. Patient Educ. Couns. **87**(2), 250–257 (2012)
35. Lu, Y., Zhang, P., Deng, S.: Exploring health-related topics in online health community using cluster analysis. In: 2013 46th Hawaii International Conference on System Sciences (HICSS), pp. 802–811. IEEE (2013)
36. Wiesner, M., Pfeifer, D.: Adapting recommender systems to the requirements of personal health record systems. In: Proceedings of the 1st ACM International Health Informatics Symposium, pp. 410–414. ACM (2010)
37. Davis, D.A., Chawla, N.V., Blumm, N., Christakis, N., Barabási, A.L.: Predicting individual disease risk based on medical history. In: Proceedings of the 17th ACM Conference On Information and Knowledge Management, pp. 769–778. ACM (2008)
38. Zhang, Y., Chen, M., Huang, D., Wu, D., Li, Y.: iDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization. Future Gener. Comput. Syst. **66**, 30–35 (2017)
39. Fillmore, C.: Frame semantics. In: Linguistics in the Morning Calm, pp. 111–137 (1982)
40. Gangemi, A.: What's in a Schema? pp. 144–182, Cambridge University Press, Cambridge (2010)