Inf Technol Manag (2009) 10:253–265
DOI 10.1007/s10799-009-0061-6

Identifying fall-related injuries: Text mining the electronic medical record

Monica Chiarini Tremblay · Donald J. Berndt · Stephen L. Luther · Philip R. Foulis · Dustin D. French

Published online: 24 November 2009

Abstract Unintentional injury due to falls is a serious and expensive health problem among the elderly. This is especially true in the Veterans Health Administration (VHA) ambulatory care setting, where nearly 40% of the male patients are 65 or older and at risk for falls. Health service researchers and clinicians can utilize VHA administrative data to identify and explore the frequency and nature of fall-related injuries (FRI) to aid in the implementation of clinical and prevention programs. Here we define administrative data as structured (coded) values that are generated as a result clinical services provided to veterans and stored in databases. However, the limitations of administrative data do not always allow for conclusive decision making, especially in areas where coding may be incomplete. This study utilizes data and text mining techniques to investigate if unstructured text-based information included in the electronic medical record can validate and enhance those records in the administrative data that should have been coded as fall-related injuries. The challenges highlighted by this study include data extraction and preparation from administrative sources and the full electronic medical records, de-indentifying the data (to assure HIPAA compliance), conducting chart reviews to construct a "gold standard" dataset, and performing both supervised and unsupervised text mining techniques in comparison with traditional medical chart review.

Keywords Healthcare informatics · Electronic medical records · Text mining · Cluster analysis · Latent semantic indexing · Veterans administration

M. C. Tremblay (✉)
Decision Sciences and Information Systems, Florida International University College of Business Administration, Miami, FL, USA
e-mail: tremblay@fiu.edu

D. J. Berndt
Information Systems and Decision Sciences, University of South Florida College of Business, Tampa, FL, USA
e-mail: dberndt@coba.usf.edu

S. L. Luther
HSR&D/RR&D Center of Excellence: Maximizing Rehabilitation Outcomes, James A. Haley Veterans Hospital, Tampa, FL, USA
e-mail: steve.luther@va.gov

P. R. Foulis
James A. Haley Veterans Hospital, Tampa, FL, USA
e-mail: philip.foulis@va.gov

D. D. French
Indianapolis VA Center of Excellence, Regenstrief Institute Inc, Indianapolis, IN, USA
e-mail: dustin.french2@va.gov

D. D. French
Division of General Internal Medicine and Geriatrics, Indiana University School of Medicine, Indianapolis, IN, USA

1 Introduction

The planning and delivery of healthcare services is an information intensive activity. The health system captures data at many points from birth and death registries, vital statistics data repositories and insurance claims data. By far the richest and complex source of healthcare information is represented by medical record entries documenting every service/encounter provided in the system. In the VHA there is a nationwide electronic medical record containing hundreds of millions of administrative and text-based records. Data captured in the electronic medical record are used for a variety of purposes, including administrative processes

and the delivery of care. Administrative data supports a variety of business processes such as insurance claims and reimbursement, as well as strategic planning. Administrative data is highly structured and more likely to be consistently applied since it is an integral part of the billing process. The structured data items include International Classification of Disease (ICD) codes for diagnoses and treatments, as well as patient characteristics. The more detailed medical record contains structured data, such as patient vital signs, laboratory results, and drug prescriptions, as well as unstructured clinical notes documenting the episodes of care. Much of the rich clinical information resides in these free text notes and is difficult to capture for use in business planning, policy formulation, government regulation, and healthcare research. The involvement of clinicians in documenting patient care and progress using the electronic medical record directly supports high quality data collection efforts and effective user interfaces for healthcare delivery. The billing requirements and care plans associated with specific treatments and procedures, with a direct financial incentive, means this data is also carefully documented. However, additional information that does not directly relate to billing or may not be critical to clinical decisions is likely to be missing or inaccurate. For instance some clinically relevant information may be embedded in textual components of the electronic medical record, accessible to clinicians at the point of care, but not easily used in aggregated form for business decisions or healthcare research. One such example is information about fall-related injuries. This information is typically recorded in various clinical notes as part of the electronic medical record, but may not be completely coded in the administrative or structured data. In these situations, text mining provides an alternative mechanism for uncovering information that can then be explicitly coded, or otherwise included in database queries that aggregate information in support of analysis and decision making.

This research focuses on using statistical text mining techniques to extract information from the electronic medical record (EMR) that can then be used to better classify those records that were not properly coded in the administrative data as fall-related injuries. Past research by healthcare investigators has indicated that the EMR can be effectively used to identify adverse events [1] or to access quality of care [1, 2].

Data and text mining of medical data is quickly gaining popularity, but is still in its infancy [3], thus the available guidance is somewhat limited. The main contribution of this paper is to describe a methodology that extracts information from the EMR in order to find and correct miscoded information in administrative data held in a different database. We demonstrate our methodology with fall-related injuries (FRI) in an outpatient setting in the VHA. Additionally, this research evaluates the feasibility and usefulness of statistical text mining techniques for other miscoded or under diagnosed cases in medicine.

This methodology includes several important and complex tasks. The first task is pre-processing, which includes selecting, cleaning and combining relevant data from administrative sources and the electronic medical records for all veterans treated for injuries and making that data available in an integrated database. The second task is de-identifying this data to assure compliance with HIPAA regulation, without losing the ability to follow the individual care of a patient and the relationship between visit dates. The third step is designing a strategy to group information from the electronic medical record related specifically to the treatment or "episode of care," with a goal of approximating which records are to be combined for a patient in order to accurately illustrate a treatment cycle.

Codes are not always correct in administrative data with the potential for false positives and false negatives [4]. In order to train models with accurate data, the fourth task is the creation of an interface which allows a registered nurse to conduct a chart review to construct a "gold standard" dataset. The nurse examines each episode of care and reads the clinical notes in the EMR in order to validate or correct the administrative data. The results are then validated by two independent coders. The final task is to create text mining models that identify incorrect or under-coded episodes of care in the administrative data.

1.1 Fall-related injuries

Unintentional injury hospitalizations due to falls in the total population and for those ages 65 and older are a serious health problem, both in the United States and throughout the world. Yet little is known about the epidemiology of these events [5–9]. Evidence suggests that falls are the costliest category of injury among older persons [6]. Among older adults, falls are the leading cause of injurious deaths and are the most common cause of nonfatal injuries and hospital admissions for trauma (CDC 2008). In 2005, 15,800 people 65 and older died from injuries related to unintentional falls; about 1.8 million people 65 and older were treated in emergency departments for nonfatal injuries from falls, and more than 433,000 of these patients were hospitalized [10–14].

In 1999, approximately 38% of male veterans were age 65 and over (compared to 13% for males in the United States). The number of "oldest old" veterans (age 85 and over) who are at highest risk of suffering injurious falls, is projected to increase dramatically from 154,000 in 1990 to 1.3 million in 2010 [15]. Due to the high incidence of falls among those aged 65 and over, it is likely that the treatment

of fall-related injuries represents a large volume of service in the Veterans Health Administration (VHA) system. Given the aging population served by the VHA healthcare system, and the high rate of unintentional injurious falls among older adults, accurate information about the incidence, prevalence and epidemiology of injurious falls is essential to clinicians, researchers and policy makers [16].

Traditionally, health services researchers and policy makers have relied on FRI documented in administrative databases to study adverse events. The structure of the data (typically designed for reimbursement purposes), the validity of the data (codes assigned correctly), and the reliability of the data (consistent application of coding practices) limit the use of these administrative data for research. Over the past decade the Veterans Health Administration has invested extensively in the implementation of an electronic medical record (EMR) system. This rich source of clinical data holds the tremendous promise of expanding the ability to conduct health services research on adverse events and other health problems. For instance, the written medical record presumably contains specific references to falls; however, these text-based notes are not easily searched. If the VHA data resources are to be fully utilized to study and monitor FRI and other adverse events, new techniques to identify and characterize injuries need to be developed and validated. One strategy that can be used to address these issues is the utilization of electronic medical record (EMR) and administrative data in the VHA healthcare information systems for decision support and knowledge management [17].

2 Methodology

Text mining is an emerging technology characterized by a set of techniques and tools which allow for the extraction of structured information from free text [18, 19]. The conceptual framework for this study is based on the application of the Cross Industry Standard Process for Data Mining (CRISP-DM) model (see Fig. 1). This process model offers a general framework for data and text mining projects, highlighting the key tasks involved (see the CRISP-DM Web page, www.crisp-dm.org). According to the CRISP-DM framework, the life cycle of a knowledge discovery project consists of six phases, but the sequence of the phases is not strictly applied. Moving back and forth between different phases is always required. The process is iterative because the choice of subsequent phases often depends on the outcome of preceding phases.

The first phase, business understanding is grounded in the importance of patient safety and the burden of FRI to the VHA. The second phase, data understanding, starts with an initial data collection effort and proceeds with activities that help the researchers become familiar with the data, to the identification of data quality and data privacy/security problems. The third phase, data preparation, involves the extraction of relevant data for a particular modeling effort, data quality assurance, and any transformations required for specific modeling techniques. Among the most important data preparation activities was a chart review to construct a "gold standard" dataset with correctly labeled FRI.

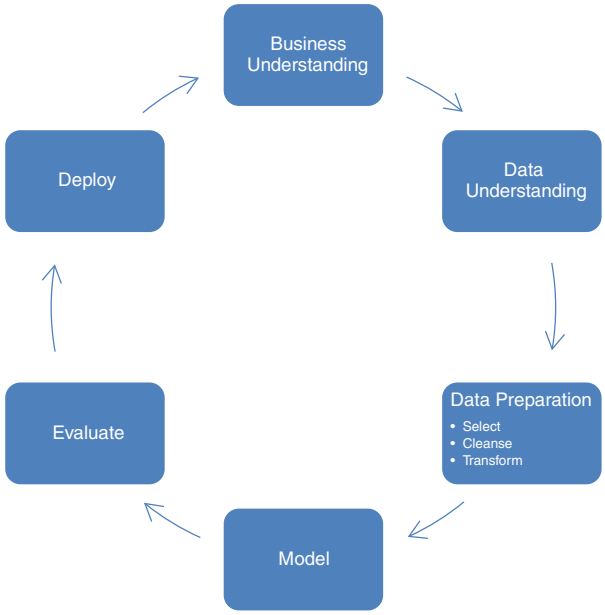

Fig. 1 CRISP-DM framework

The fourth phase, data modeling, is the central focus of any knowledge discovery effort and consists of the construction of models using a variety of techniques. In this research, two different term weighting approaches are applied to events or episodes of care in the electronic medical records: entropy and information gain. The terms created by those weighting schemes are then used in two different approaches: clustering (an unsupervised learning technique) and logistic regression (a supervised learning technique). Our goal is to assess the potential predictive power of these textual descriptions in identifying FRIs. If terms found in the electronic medical records are indicative of FRI, we would expect the formation of clusters that are comprised primarily of FRIs. Similarly, those terms should be useful in the classification of FRIs. The fifth step, evaluation, is conducted for both weighting approaches. The cluster models are evaluated by exploring the formed clusters based on occurrences of FRIs (the goal is to see clusters formed with mostly one outcome, such as FRIs). The text mining terms are also used to form predictive logistic regression models that are evaluated by classification matrices and overall accuracy, sensitivity, and specificity.

Though we have not conducted the final step, deployment, we discuss two possible strategies for incorporation of the results into the VHA medical records system: running the algorithms on historical data to better support health policy research and analysis and/or supporting human coders by using the algorithms to generate customized drop-down lists and other navigation aids to improve the coding process. The results could be embedded in decision support systems, prompting clinicians to assign the correct injury (or E-code) based on the electronic medical records just entered, or used to more completely code existing electronic medical records. Correctly coded data can then aid the VHA in identifying the frequency and nature of fall-related injuries in order to better organize and implement prevention strategies.

2.1 Business understanding

External Cause of Injury codes (commonly called E-codes) are a supplemental code for use with the International Classification of Diseases (ICD) that provide a systematic way to classify diagnostic information that doctors, nurses, and other health care providers have entered into the medical record.[1] E-codes should provide an efficient way to identify fall-related injuries (FRIs) from the administrative data; but this is contingent on consistent and accurate use by providers and coders. Studies are not congruent on the use of E-codes. One study at the state level found that E-codes assigned by professional coders in hospitals can be very accurate [20]. Other national studies have found considerable variation with higher rates in states with mandates for recording E-codes [21, 22]. In the VA, E-codes are recorded by clinicians upon completion of clinical interactions. Given the workload of VA clinicians and the fact that E-codes are not required to be recorded, FRIs are likely to be under-coded. One of the goals of this research is to investigate the ability of text mining to identify FRIs based on clinical notes in the electronic medical record (EMR).

2.2 Data understanding

Data understanding in the context of this study can be described as the clinical or treatment information about FRI available in the EMR. The results of the text-mining analysis will improve our ability to identify FRI from secondary data through the systematic analysis of text-based data. The descriptive phrases and fragments of EMRs identified during text mining represent a set of variables that are predictive of FRI. These variables may represent factors known to be associated with FRI, but typically not available in administrative databases. Several studies have shown that the risk of falling increases dramatically as the number of risk factors increases, ranging from a probability of 0.27 with zero or one risk factor to 0.78 for those with four or more risk factors [23–25]. The susceptibility to injury in older adults also stems from a high prevalence of co-morbid diseases (e.g. osteoporosis, sarcopenia) as well as age-related decline (e.g. slowed reflexes), which can make even a relatively mild fall very dangerous. A multitude of risk factors for falling have been reported, including lower extremity weakness, gait and balance disorders, previous falls, functional impairment, visual deficits, cognitive impairment, depression, polypharmacy, and stroke [26–29]. Many of these risk factors are not reliably coded in administrative data sets, yet providers may document these issues within the textual clinical notes in the electronic medical record.

2.3 Data preparation

Data cleaning and preprocessing almost always accounts for a large share of any data mining efforts. As Fayyad (1996, pg.30) points out:

> "In practice, a large portion of the applications effort can go into properly formulating the problem (asking the right question) rather than optimizing the algorithmic details of a particular data mining method."

In this study there was a large amount of available data, which is normally an ideal condition, but one which can introduce many challenges. The processes of understanding which data are important, what techniques to use to combine data from two very different sources, understanding how to manipulate and process very large text fields (such as electronic medical records), and complying with both HIPAA (Health Insurance Portability and Accountability Act of 1996, which establishes regulations for the use and disclosure of a patient's medical record or payment history) and VHA privacy standards are among the challenges faced in this study. In addition, the availability of many rich textual fields in the VHA data made it crucial to fully understand how to best select keywords and concepts with the purpose of building predictive models. Figure 2 outlines our data preparation process.

2.3.1 Data selection

The Veteran's Administration Medical SAS Inpatient (Event) Database[2] was searched to identify all outpatient

[1] http://edc.org/buildingsafecommunities/buildbridges/bb2.2/ECODES.html, 2008.

[2] The VHA Medical SAS Datasets are national administrative data for VHA-provided health care utilized primarily by veterans. The

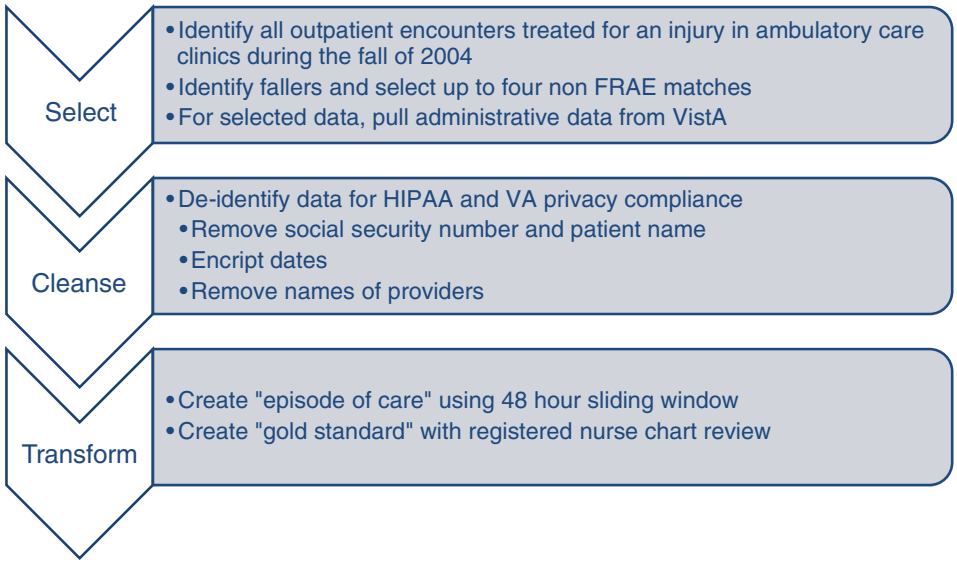

Fig. 2 Data preparation process

encounters treated for an injury in the James A. Haley Veterans Administration Hospital (corporate) ambulatory care clinics during the fall of 2004. The initial data set was further refined by excluding encounters for spinal cord injury, poisoning and iatrogenic injuries. From these data a subset of unique patients who were treated for falls were identified. For the purposes of this study, we define the situation in which a veteran seeks treatment for a fall-related injury in the ambulatory care setting as a fall-related ambulatory event (FRAE). This FRAE definition provides an efficient method to estimate the occurrence of FRI in lieu of expensive primary data collection efforts. Event level data was collapsed into episodes of care to identify unique fall events.

Because it was anticipated that FRAEs would be under reported, we sought to develop a pool of ambulatory care patients from the initial data set who had received care for an injury as the primary reason for the visit but were not identified as being fallers. Therefore, we matched cases identified as being fallers (at least one FRAE) in the previous data step with the pool based on gender, age and primary injury code. We identified up to three or four matches for each fall case. As a next step, all electronic and administrative records and clinical EMRs were obtained for the patients in the analytic data set by searching VistA (Veterans Health Information System and Technology Architecture). VistA is an enterprise-wide information system built around an electronic health record.

2.3.2 *Data cleansing*

The next step in preprocessing involved de-identifying this data and loading it into a relational database. The social security number and patient name were removed and a unique number corresponding to an individual patient was assigned. The unique number was maintained throughout the database allowing us to follow the care of an individual de-identified patient. The policies designated to assure that research is compliant with HIPAA regulations also required that dates of service be encrypted in the data used for analysis. The dates of service were encrypted by generating a random number and assigning it to all dates for each unique patient. A number of days equal to this random number were then subtracted from each date. This process ensured that the original dates were eliminated from the data, while the relationship between dates for a particular patient was maintained. Thus, the exact date was not available for the researchers and patient confidentiality was maintained. Finally, all variables that contained the names of providers were eliminated from the data.

2.3.3 *Data transformation*

A patient identifier and date were used to link the ambulatory events data to the text based medical records (unfortunately there was not a unique identifier). Multiple encounters may appear for each patient in a selected time period. Thus, a patient could have a series of electronic

Footnote 2 continued
datasets are provided in SAS® format by fiscal year (Oct. 1 - Sept. 30). These data are extracted from the National Patient Care Database (NPCD). For more see http://www.virec.research.va.gov/datasources name/Medical-SAS-Datasets/SAS.htm.

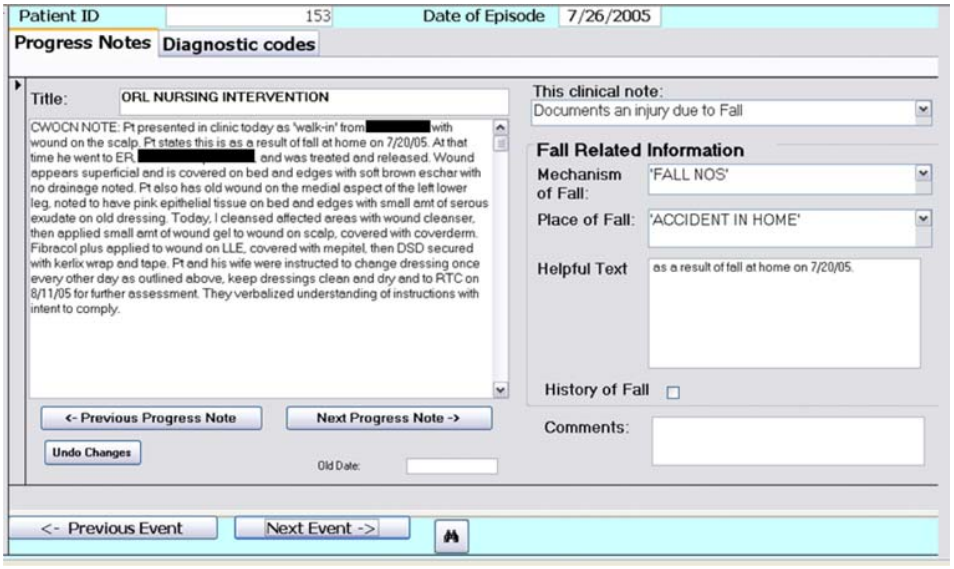

Fig. 3 Benchmarking interface

medical records for events that were in fact related, describing the normal process of a patient proceeding through the outpatient facility. Deciding how to group these records, thereby creating an "episode of care" is crucial. A brute force approach is to read each record and to manually combine these records to provide accurate representations with which to train a model. The obvious problem with this approach is that it is extremely time consuming and would not be feasible for regional or national data. Another approach is to create a "sliding window," allowing for some overlap and deciding on a time frame in order to automate this process. Our initial model used a simple approach of defining the episode of care as all the notes and data collected during a 48 h window. We believe this window length captures most relevant information for two reasons: first, most clinical notes are recorded immediately after the encounter and second, if notes are not completed immediately, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) dictates they be completed within 48 h.

As is commonly done for medical expert systems [30, 31], and in particular medical text processing and information retrieval studies [32], we created an expert annotated version of the dataset, or "gold standard." Since we are unsure whether the administrative data is correctly coded,[3] we devised a GUI interface that allowed a registered nurse to conduct a chart review (see Fig. 3 for an example). The registered nurse was hired to read through the 48 h episodes of care and to code whether the clinical note indicated a fall-related injury, as well as the mechanism of fall, the place the fall occurred and any helpful fragments found in the text. This additional information is captured for future research endeavors, such as increasing the sophistication of our models to predict detailed codes such as mechanisms of falls.

The nurse received training prior to beginning the chart review, and intra-rater and inter-rater reliability was established with two other independent coders (the agreement percentage was 95%). This dataset, which consisted of approximately 1500 events, was utilized for our initial text mining analysis.

Chart reviews confirmed 72% of the original cases identified through administrative data and found an additional 18% of cases not identified through administrative data alone.

3 Modeling

Figure 4 illustrates our modeling and evaluation approach. First, we label each created episode of care as FRI or non-FRI based on the chart review. We apply statistical text mining using two different weighting schemes, and create two datasets for analysis: one dataset containing the terms from an entropy weighting scheme (an unsupervised weighting technique), and one with the terms from information gain (a supervised weighting technique). We split each dataset into a training dataset (80% of the data) and a validation dataset (20% of the data). We then use the terms created by both weighting schemes in two separate tasks: 1) unsupervised learning (k-means clustering) and 2) supervised learning (classification models using logistic regression).

[3] An E-code should indicate a fall related injury.

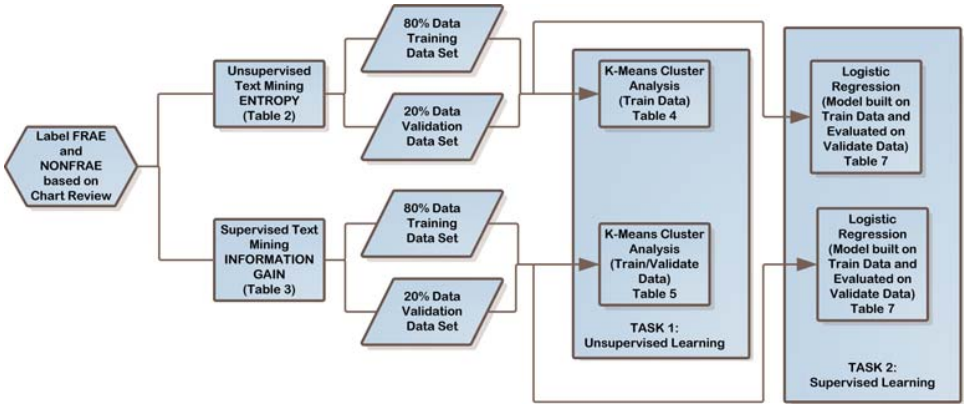

Fig. 4 Modeling process

3.1 Text mining

There are three fundamental approaches to text mining: simple keyword or regular expression matching, natural language processing, or machine learning algorithms. Simple keyword or string matching approaches are typically not powerful enough for extracting information from clinical notes. Natural language processing (NLP) approaches are deductive in nature, using clinical theory to construct controlled vocabularies and ontologies that underpin the text processing, with additional components to handle negation and temporal relationships. These approaches apply both syntactic and semantic rules to represent or "comprehend" clinical concepts in documents. The vocabularies, standardized coding systems, and concept ontologies necessary for NLP approaches require significant development investments. The medical domain is one of the few fields to make decades-long investments in these fundamental language processing resources and achieve reasonable results in a variety of applications [33–35]. NLP approaches can be used for classification, as well as more targeted information extraction, such as finding a specific type of laboratory result. In most cases these tasks require some degree of customization or enhancements to the controlled vocabularies or rule sets. In contrast, machine learning algorithms are inductive, data-driven approaches to text mining that do not rely on controlled vocabularies or ontologies. These algorithms take a more statistical approach, calculating word frequencies and term weights to discriminate between or group documents by similarity measures. Statistical text mining algorithms are well suited to classification tasks like the labeling of fall-related injuries as part of this research. These methods can be used without much customization and are free of any biases that may be implicitly embedded in the relevant portions of handcrafted vocabularies or coding systems. This paper explores the use of such machine learning algorithms to extract structured information from clinical notes within a flexible, fairly lightweight application environment.

Many of the machine learning algorithms used for text mining "count" occurrences of words or phrases in documents. To simplify this, most algorithms generally remove specified words (stop list) or keep specified words (start list). Words that have a common root are stemmed, and common words are removed since they have little power in discriminating documents [36, 37]. A term-by-document frequency matrix is built and serves as the foundation for analysis of the document collection. To improve performance, entries can be adjusted by utilizing weighting functions for certain words (e.g., infrequent words may be weighed more heavily, along with words that are highly correlated to a target variable).

An important decision is selecting a weighting scheme that can help emphasize discrimination between document groups. The original frequencies in the term-document frequency matrix are transformed to the "expected" frequency as follows: $\hat{a}_{ij} = L_{ij}G_i$, where L_{ij} is the frequency weight and G_{ij} is the term weight. There are several options for frequency weights, but because most common raw frequencies have values of 0 or 1, the default log transformation appears to have no effect [36]. In this experiment, log frequency weighting was used where $L_{ij} = \log_2(a_{ij} + 1)$, and a_{ij} is the frequency with which term i appears in document j (note that for $a_{ij} = 1$, $L_{ij} = 1$ and when $a_{ij} = 0$, $L_{ij} = 0$).

The choice of *term importance* weightings (G_{ij}) can have a large impact as well. Term importance weightings give higher weights to those terms that are more important than others. Research has shown that good results are often obtained using entropy or inverse document frequency [38, 39]. The weightings can also be target-based (in our case whether a given record was fall-related or not), which requires training data that is correctly labeled. This is one of the primary reasons for the chart review.

The resulting weighted term-by-document frequency matrix can grow quite large. Additionally, most of the terms are not used in all the records (thus many columns will have a value of 0), yielding a sparse matrix and a heavy computational burden. Therefore, it is typically advantageous to reduce the dimensionality of this matrix. Latent semantic indexing (LSI) reduces dimensionality by using singular value decomposition (SVD). LSI is a technique that transforms the large matrix into a much lower dimensional form [38, 40]. SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. Singular value decomposition is closely related to eigenvector decomposition. Similar to factor analysis, the frequency matrix is decomposed into eigenvalues and eigenvectors that create linearly independent components of the data. The smaller components can be ignored and the similarity between two documents can be determined by the values of the remaining factors. The result can be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two documents corresponds to their estimated similarity [38]. For more detailed coverage of these issues, see [38, 40].

These algorithms were applied to the clinical notes found in the electronic medical record using SAS Enterprise Miner. We utilized the tool's automated stemming of terms (for example BIG: BIG, BIGGER, BIGGEST), as well as modifying the initial synonym lists and stop lists by combining the terms that were synonyms and adding unnecessary words to the stop list. For example, the word "outpatient" appears often because these are outpatient records and is therefore not a very useful discriminatory word.

We explore two term importance weightings: entropy (unsupervised weighting) and information gain (supervised weighting) and perform two tasks using the terms identified by each weighting scheme: cluster analysis (an unsupervised technique), and logistic regression (a supervised technique). We split the datasets into training and validation datasets (see Table 1) to evaluate the models on data that was not used to train the models (and to avoid overfitting). The goal was to classify each encounter as a fall or non-fall, using a derived weighting scheme that emphasized discrimination between notes associated with FRAE and notes not associated with FRAE.

Table 1 Dataset description

Data	Non-FRAE number (%)	FRAE number (%)	Total episodes number (% of total data)
All	1464 (68%)	693 (32%)	2157 (100%)
Train	1170 (68%)	553 (32%)	1723 (80%)
Validate	294 (68%)	140 (32%)	434 (20%)

3.2 Weighting techniques

We apply two term importance weighting approaches that will generate SVDs to be used for cluster analysis and logistic regression (LR). The unsupervised weighting technique allows us to investigate if there is sufficient difference between the FRAE EMR and the non-FRAE EMR that even given an unsupervised weighting scheme, separate clusters form for each category, and the SVDs will create models using LR that are able to classify FRAE/non-FRAE. With the supervised approach, we expect that the cluster algorithms will generate clusters that have stronger concentrations of either FRAE or non-FRAE, and that the LR models will be more accurate since the importance weightings are created with knowledge of whether the event was a FRAE or not. For both approaches, we hypothesize that it is possible to both identify clusters of terms that are indicative of a FRAE and to classify events as FRAEs. Both approaches provide evidence that the electronic medical record text will have predictive power in identifying fall-related injuries.

3.2.1 Entropy: Unsupervised term importance weighting

Entropy is a concept from communication theory [41] and is a measure of information content (i.e., disorder). Entropy gives high weights to terms that are infrequent in all the data, but frequent in a few documents [36]. We select the entropy technique because it does not take into account our target variable (FRAE/non-FRAE) to create the weighted term-by-document frequency. The goal of this step is to see if it is possible to identify terms that are indicative of a FRAE. This would indicate that the fall-related medical records are sufficiently different from other documents (keeping in mind that all the documents are injury related). Table 2 shows a fragment of the weighting scheme using entropy for unsupervised clustering.

Table 2 Partial example of frequency table with weights using entropy

Term	Freq	# Documents	Keep	Weight	Role
+Pain	1639	510	Y	0.151	Noun
+He	1314	364	Y	0.215	Pron
Active	1090	136	Y	0.343	Prop
Pain	893	407	Y	0.184	Prop
+Fall	461	348	Y	0.167	Verb
+Will	404	225	Y	0.249	Aux
+List	330	179	Y	0.277	Verb
+Tablet	301	69	Y	0.430	Noun
Medications	293	157	Y	0.284	Prop

3.2.2 Supervised term importance weighting

We now utilize a similar technique, but select an information gain weighting scheme, which considers the target variable indicating whether the document was a FRAE. Information gain is also an entropy reduction technique and indicates how well the term, or the absence of that term, predicts the category (by calculating a reduction in entropy). Table 3 illustrates a partial example of the frequency table for this weighting scheme.

3.3 Cluster analysis

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison with one another, but are dissimilar to objects in other clusters [42]. Clustering classifies the data into groups based on measures of distance or similarity, and is an excellent way to group documents into thematic categories [43]. Clustering is an unsupervised technique in that it does not take into account whether the events are FRAE or non-FRAE. We cluster based on the terms from the two term importance weighting approaches discussed above. We hypothesize that it is possible to identify clusters of terms that are indicative of a FRAE, for both weighting techniques, though we expect the information gain to form stronger clusters.

Table 3 Partial example of frequency table with weights using information gain

Term	Freq	# Documents	Keep	Weight	Role
+Fall	880	682	Y	1.000	Verb
+Fall	550	394	Y	0.460	Noun
Xray	202	152	Y	0.251	Prop
+Abrasion	176	113	Y	0.183	Noun
+Month	354	305	Y	0.170	Noun
+He	4503	1132	Y	0.161	Pron
+Injury	328	268	Y	0.161	Noun
+Continue	467	373	Y	0.141	Verb
+Treatment	286	222	Y	0.135	Noun

3.3.1 Clustering using an entropy weighting scheme

Table 4 shows the clusters identified using k-means clustering on the SVD terms created using an entropy weighting scheme on 1,721 randomly sampled records (where 32% of the records are FRAE). The columns Flag = Y and Flag = N categorize what percentage of the records in these clusters that were either FRAEs, or not FRAEs, based on the chart review. None of the clusters has a majority of fall events. An unsupervised weighting scheme, coupled with an unsupervised data mining approach, did not produce good results in identifying fall

Table 4 Clusters identified from terms in the electronic medical records (using entropy)

Cluster	Flag = Y number (%)	Flag = N number (%)	Descriptive terms	Total number (%)
A	304 (35%)	574 (65%)	Active, swelling, history, +deny, +week, +pain, +male, +fracture, s/p, +year, +foot, old, ago, male, left, +do, +patient, +present, but, +fall, +see, +shoulder, when, +fall, +not	878 (51%)
B	61 (22%)	214 (78%)	Ref, +small, tissue, dated, today, +edge, pt, +area, will, +apply, cwocn note, cwocn, +bed, +heal, skin, drainage, +clinic, note, +note, +dressing, +wound, +continue, with, +leg, home	275 (16%)
C	80 (41%)	117 (59%)	Temperature, ambulatory, vista, height, triage, bp, patient, +obtain, chief, kg, bmi, cm, vitals, recorded, recent recorded, recent, patient vitals, weight, pulse, nka, nursing triage assessment, patient chief complaint, ambulatory history, nka chief complaint, nursing	197 (11%)
D	73 (29%)	179 (71%)	+Day, recent, +note, +complaint, =-=-dshi disposition report-=-=, =-=-dshi, +list, patient, telecare, length, evaluation/management, ending, author, physician phone consultation, +assign, starting, area, encounter, caller, duration, +minute, evaluation, +comment, health, +type	252 (15%)
E	35 (29%)	84 (71%)	Other, va, nka, alert, signs, visit, 's, scale, +answer, is, screen, +provider, reason, +orient, pain, +relate, +visit, lb, mobility, impairment/, is other provider, treating pain, treating, cognitive screen, x3	119 (7%)
Unlabeled	2 (100%)	0		2
Total	1170 (68%)	553 (32%)		1723

related events. Since the results were not encouraging, we did not bother to replicate the clustering results on validation data, which we would expect to have a similar performance.

3.3.2 *Clustering using an information gain weighting scheme*

We take the same approach and use k-means clustering on the SVD terms created using the information gain weighting scheme on 1,721 randomly sampled records (where 32% of the records are FRAE). As is illustrated by Table 5, the much better results are achieved. Cluster A consists of 90% of FRAEs, indicating the majority of the events for this cluster are FRAEs. We replicate on our validation data (results shown in Table 6), achieving similar results. Cluster A contains 86% of the FRAEs. Again, this requires correctly labeled training data. The chart review conducted as part of data preprocessing provided much more reliable training data than was available by automated extraction.

3.4 Predictive modeling using logistic regression

Like linear regression, logistic regression is part of a category of statistical models called generalized linear models. Logistic regression, however, allows for the prediction of a discrete outcome, such as FRAE and non-FRAE from a set of variables. In this case we compare the results for two predictive models using LR: one using the terms and SVDs from the entropy weighting scheme, and one using the terms and SVDs items from the information gain weighting scheme. Table 7 lists the resulting sensitivity and specificity as well as the classification matrix for the validation data (we do not report these for the training data, because validation data better describes true predictive capability of the model). As expected, the model that used the terms and SVDs from the information gain out-performed the entropy weighting scheme. It is important to note that sensitivity (83%) and specificity (93%) for the unsupervised weighting technique indicate a model with strong predictive capability. This is encouraging since, as was previously noted, correctly labeled data is often difficult to obtain.

3.5 Further analysis

The terms and SVD factors that were selected for inclusion in the logistic regression model should reflect a plausible basis for discriminating between non-fall and fall-related injuries. Table 8 shows the individual rollup terms and SVD factors that were found most predictive. The individual terms are quite relevant including "falling" and "fx" (the abbreviation for fracture). The SVD terms also

Table 5 Resulting clusters using information gain

Cluster	Flag = Y number (%)	Flag = N number (%)	Descriptive terms	Total number (%)
A	458 (90%)	53 (10%)	+Fall, +fall, +rib, when, +arm, +side, home, left, ago, +do, +deny, er, history, +state, +knee, pulse, pt, complaint, on, +list, +shoulder, right, +pain, vista, +not	511 (30%)
B	69 (8%)	792 (92%)	+Fall, +fall, +abrasion, +knee, back, past, regular, +leg, mental, +vital, +injury, +month, during, +hand, +tablet, active, medications, some, +wrist, +orient	861 (50%)
Unlabeled	26 (7%)	325 (93%)		351 (20%)
Total	1170	553		1723

Table 6 Replication of clusters using information gain weightings

Cluster	Flag = Y number (%)	Flag = N number (%)	Descriptive terms	Total number (%)
A	118 (86%)	20 (14%)	+Fall, +fall, +rib, when, +arm, +side, home, left, ago, +do, +deny, er, history, +state, +knee, pulse, pt, complaint, on, +list, +shoulder, right, +pain, vista, +not	138 (32%)
B	15 (93%)	190 (7%)	+Fall, +fall, +abrasion, +knee, back, past, regular, +leg, mental, +vital, +injury, +month, during, +hand, +tablet, active, medications, some, +wrist, +orient	205 (47%)
Unlabeled	7 (8%)	84 (92%)		91 (21%)
Total	294	140		434

Table 7 Hit rates for LR on validation data

	Entropy	Information gain
Accuracy	90%	91%
Sensitivity	83%	89%
Specificity	93%	93%
True positive	116	124
False positive	20	22
True negative	274	272
False negative	24	16

Table 8 Individual rollup terms and SVD factors that were found most predictive

Rollup terms	SVD 4	SVD 1
Change pysch	Trip	Fall
fx	Sandbag	Chief complaint pt
Falling	Rug	Complaint pt
	Electrical cord	Out of
	Assistive	Alert
	Assitive device	Today
	Device	Out
	Cord	In
	Electrical	Come
	Wooden	Outstretch
	Through	Last
	Through rug	Time

summarize sets of important terms. For instance, SVD 4 heavily weights "trip," "rug," "assistive device" (such as walkers or canes), and what looks like a misspelling of "throw rug." SVD 1 emphasizes the obvious term "fall" along with "chief complaint." Both the individual terms and SVD factors do seem to combine to provide a reasonable model for identifying fall-related injuries.

3.6 Limitations

If we more closely investigate those records that were correctly and incorrectly classified by our supervised clustering algorithm we can investigate some of the limitations of this technique. Table 9 illustrates four example clinical notes, comparing the text mining results to the chart review. The first two records were incorrectly identified by the text mining algorithm. The first record contains words such as hip, pain, and knee that might also be commonly found in a FRAE, but it is clearly not a FRAE record (with no mention of a fall). The second record, however, is clearly a miscoded FRAE. In this case a fall occurred, but without a definitive injury, it mentions falling briefly, but mainly discusses other issues. These types of incorrectly coded records may cause trouble during the model training process and evaluation, highlighting the need for chart reviews and the need for further case investigation in order to improve the text mining algorithms.

4 Conclusions

In this paper we utilize data and text mining techniques to investigate if unstructured text-based information included in the electronic medical record can be used to validate and enhance incorrectly coded records in administrative data. In particular we explored the ability of text mining to identify FRIs based on clinical notes in the EMR.

Table 9 Comparison to chart review

Record	EMR	Text mining	Chart review
1	Patient here for pain in the right knee. Denies any trauma but states she was just going to sit down on the couch, while holding a small child, and felt this burning type sudden pain in the medial knee area. Has been about the same since then. C/O some pain into the right hip and now in the medial knee area again. Straightening the knee helps. She states has numbness in the three toes (last ones)	Fall	No fall
2	This RN was called to Physical Therapy department to assess resident. As informed by staff resident had 'fainting episode'. Resident was placed in w/c and had vital signs taken. Awake, alert and oriented x3,stated:'This happens to me all the time; that's why I fall. I get dizzy and everything goes dark'. Taken to room and placed in bed; EKG performed. ARNP also aware. Resident was able to eat and tolerate lunch. No further complaints voiced	No fall	Fall
3	Patient called for appt today. States he tripped over a bag of leaves at night and fell and hurt his right knee-had big cut on the right knee-not healing well-with yellow discharge	Fall	Fall
4	Patient out of town for appt. Rx issued today. 2. Patient states taking Medication as directed for chronic pain and states Rx is effective. 3. Understands directions and use 4. RTC in four weeks. 5. Instructions not to, drink alcohol, drive or operate mechanical equipment while taking narcotics given, verbalizes understanding	No fall	No fall

We described several challenges in selecting, extracting, and transforming both the data from administrative sources and the full electronic medical records, and de-indentifying the data (to assure HIPAA compliance). We performed both supervised and unsupervised text mining techniques as well as supervised and unsupervised data mining techniques to ascertain whether "good enough" results could be obtained even if we were not always certain of the classification of the EMR note.

The initial results are encouraging and demonstrate that conducting text mining on electronic medical records will be helpful in correctly identifying fall-related injuries. In fact, it appears that text mining alone may be useful for identifying fall-related injuries, at least as a binary classification problem. For the text mining algorithm, we investigate two weighting schemes. We then use the terms generated by both weighting schemes to conduct two forms of analysis: clustering, an unsupervised technique, and logistic regression, a supervised technique. As expected, information gain outperforms entropy for both clustering and LR-based classification, since it considers whether the record was a FRI as it assigns weights. Using an entropy weighting technique and then clustering on those terms did not produce viable results. However, it is encouraging to note that even entropy performs well for the classification task, since in many cases correctly labeled data is not available. In fact, our expert annotated version of the dataset found a good number of FRIs that were not coded as such in the administrative data.

4.1 Future research

Our initial models only use the terms extracted by the text mining algorithms. Yet, the data available from the VHA has many other fields that can be used for predictive modeling. Future work will include more complete attribute selection from the abundant structured and unstructured data available. Accurate and extensive feature selection and preparation should improve the predictability of such data mining algorithms [44–46]. Additionally, more precise "episode of care" groupings of the electronic medical records would undoubtedly enhance the performance of the predictive models. Thus, the results of these current models are preliminary and should improve as our data preprocessing techniques evolve.

Future research will also focus on predicting more detailed E-codes that describe the type of fall-related injury, the location of the event, and other information useful in better understanding these health challenges. More importantly, this is an example of using machine learning approaches for text mining to extract useful information from clinical notes that can be used in database queries and to create summary reports for business analysis and health research.

Once more research is conducted and more robust models are constructed, the resulting models can be embedded in decision support tools, with the possibility of automatic prompting for nurses or clinicians entering data in order to assign a suggested E-code based on the electronic medical records being written. Another possibility is to use these data mining techniques to post-process medical records and add structured codes in a fully or semi-automated manner. The results can also be used as means of auditing incorrect use of E-codes in administrative data for reporting purposes. Correctly coded data can then aid the VHA in identifying the frequency and nature of fall-related injuries in order to implement prevention programs and minimize the cost and adverse effects of falls.

Acknowledgments The authors acknowledge research support of resources and use of facilities provided by the James A. Haley Veterans' Hospital in Tampa, Florida.

References

1. Thomas EJ, Studdert DM, Brennan TA (2002) The reliability of medical record review for estimating adverse event rates. Ann Intern Med 136(11):812–816
2. Baker DW et al (2007) Automated review of electronic health records to assess quality of care for outpatients with heart failure. Ann Intern Med 146(4):270–277
3. Lee IN, Liao SC, Embrechts M (2000) Data mining techniques applied to medical information. Med Inform Internet Med 25(2):81–102
4. Hunt P et al (2007) Completeness and accuracy of international classification of disease (ICD) external cause of injury codes in emergency department electronic data. Inj Prev 13(6):422–425
5. Kannus P et al (1999) Fall-induced injuries and deaths among older adults. JAMA 281(20):1895–1899
6. Rizzo JA et al (1998) Med Care 36(8):1174–1188
7. Scuffham P, Chaplin S, Legood R (2003) Incidence and costs of unintentional falls in older people in the United Kingdom. J Epidemiol Community Health 57(9):740–744
8. Koski K et al (1998) Risk factors for major injurious falls among the home-dwelling elderly by functional abilities. Gerontology 44:232–238
9. Cesari M et al (2002) Prevalence and risk factors for falls in an older community-dwelling population. J Gerontol A Biol Sci Med Sci 57:722–726
10. Nevitt MC, Cummings SR, Hudes ES (1991) Risk factors for injurious falls: a prospective study. J Gerontol 46:M164–M170
11. Nevitt MC, Cummings SR, Kidd S (1989) Risk factors for recurrent nonsyncopal falls: a prospective study. JAMA 261:2663–2668
12. Rubenstein L, Joephson K (2002) The epidemiology of falls and syncope. In: Kenny RA, Oshea D (eds) Falls and syncope in elderly patients. Clinics in Geriatric Medicine, pp 141–158
13. Tinetti M, Speechley M, Ginter S (1998) Risk factors for falls among elderly persons living in the community. N Eng J Med 319:1703–1707

14. Jager T et al (2000) Traumatic brain injuries evaluated in U.S. emergency departments, 1992–1994. Acad Emerg Med 7(2): 134–140
15. Klein R, Stockford D (2000) The changing veteran population: 1999–2020. Office of the DAS for Program and Data Analyses
16. Luther S et al (2005) Fall-related ambulatory care services in the veterans administration healthcare system. Aging Clin Exp Res 17(5):412–418
17. Kraft MR, Desouza KC, Androwich I (2003) Data mining in healthcare information systems: case study of a VeteransÂ' administration spinal cord injury population. In: HICCS, Hawaii
18. Feldman R, Dagan I (1995) Knowledge discovery in textual databases (KDT). In: Proceeding of 1st international conference on knowledge discovery (KDD-95)
19. Loh S, Oliveira JPMD, Gameiro MA (2003) Knowledge discovery in texts for constructing decision support systems. Appl Intell 18:357–366
20. Ribbeck BM, Runge JW, Thomason M (1992) Injury surveillance: a method for recording e codes for injured emergency department patients. Ann Emerg Med 21:37–40
21. Coben J et al (2001) Completeness of cause of injury coding in healthcare administrative databases in the United States. Inj Prev 12(3):199–201
22. Lawrence B et al (2007) Issues in using state hospital discharge data in injury control research and surveillance. Accid Anal Prev 39(2):319–325
23. The American Geriatrics Society (2001) B.G.s.a.A.A.o.O.S.-p.o.F.P., Guideline for the prevention of falls in older persons. J Am Geriatr Assoc 49(5):664–672
24. Nguyen TV, Eisman JA, Kelly PJ, Sambrook PN (1996) Risk factors for osteoporotic fractures in elderly men. Am J Epidemiol 144(3):255–263
25. Kraft MR, Desouza KC, Androwich I (2003) Data mining in healthcare information systems: case study of a Veterans' administration spinal cord injury population. In: Proceedings of the 36th Hawaii international conference on system sciences, Hawaii
26. Rubenstein LZ, Josephson KR, Robbins AS (1994) Falls in the nursing home. Ann Intern Med 121(6):442–451
27. Rubenstein LZ, Powers CM, MacLean CH (2001) Quality indicators for the management and prevention of falls and mobility problems in vulnerable elders. Ann Intern Med 135(8, Part 2): 686–693
28. Nevitt MC (1997) Falls in the elderly: risk factors and prevention. In: ed. Masdeu JC SL, Wolfson L (eds) Gait disorders in aging. Lippincott-Raven, Philadelphia
29. Yates JS et al (2002) Falls in community-dwelling stroke survivors: an accumulated impairments model. J Rehabil Res Dev 39:385–393
30. Evans DA, Patel VL (1992) Advanced models of cognition for medical training and practice: proceedings of the NATO advanced research workshop on advanced models of cognition for medical training and practice, held at Il Ciocco, Barga, Italy, June 19–22, 1991. Springer
31. Stead WW et al (1994) Designing medical informatics research and library-resource projects to increase what is learned. J Am Med Inform Assoc 1(1):28–33
32. Hripcsak G, Rothschild AS (2005) Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 12(3):296–298
33. Ware H, Mullett CJ, Jagannathan V (2009) Natural language processing framework to assess clinical conditions. J Am Med Inform Assoc 16(4):585–589
34. Brown SHE et al (2008) eQuality for all: extending automated quality measurement of free text clinical narratives. In: AMIA 2008, Washington DC
35. Unified Medical Language System. Available from: http://www.nlm.nih.gov/research/umls
36. Woodfield T (2003) Text mining using SAS Software course notes
37. Wei C-P, Yang CC, Lin C-M (2008) A latent semantic indexing-based approach to multilingual document clustering. Decis Support Syst 45(3):606–620
38. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
39. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
40. Berry MW, Browne M (1999) Understanding search engines: mathematical modeling and text retrieval. P.S.f.I.a.A. Mathematics
41. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:398–403
42. Han J, Kamber M (2001) Data mining: concepts and techniques. The Morgan Kaufmann Series in Data Management. M.K. Publishers. San Diego,
43. Spangler S, Kreulen JT (2008) Mining the talk: unlocking the business value in unstructured information. IBM Press/Pearson plc, Upper Saddle River, xix, 217 pp
44. Dash M, Liu H, Yao J (1997) Dimensionality reduction of unsupervised data. In: 9th International conference on tools with artificial intelligence. IEEE Computer Society, Washington DC, New Port Beach, CA
45. Tremblay MC, Berndt DJ, Studnicki J (2006) Feature selection for predicting surgical outcomes. In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06)
46. Barbara D et al (1997) The new jersey data reduction report. IEEE Data Eng Bull 20(4):3–45