Retrieving lost information from textual databases: rediscovering expeditions from an animal specimen database

Marieke van Erp
Dept. of Language and Information Sciences
Tilburg University, P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands
M.G.J.vanErp@uvt.nl

Abstract

Importing large amounts of data into databases does not always go without the loss of important information. In this work, methods are presented that aim to rediscover this information by inferring it from the information that is available in the database. From and animal specimen database, the information to which expedition an animal that was found belongs is rediscovered. While the work is in an early stage, the obtained results are promising, and prove that it is possible to rediscover expedition information from the database.

1 Introduction

Databases made up of textual material tend to contain a wealth of information that remains unexplored with simple keyword-based search. Maintainers of the databases are often not aware of the possibilities offered by text mining methods to discover hidden information to enrich the basic data. In this work several machine learning methods are explored to investigate whether 'hidden information' can be extracted from an animal specimen database belonging to the Dutch National Museum for Natural History, Naturalis[1]. The database is a combination of information about objects in the museum collection from handwritten data sources in the museum, such as journal-like entries that are kept by biologists while collecting animal or plant specimens on expedition and tables that link the journal entries to the museum register. What is not preserved in the transition from the written sources to the database is the name of the expedition druing which an animal specimen was found.

By expedition, the following event is implied: a group of biologists went on expedition together in a country during a certain time period. Entries in the database that belong to this expedition can be collected by one or a subset of the participating biologists. For researchers at the natural history museum it would be helpful to have access to expedition information in their database, as for biodiversity research they sometimes need overviews of expeditions. It may also help further enrichment of the database and cleansing, because if the expedition information is available, missing information in certain fields, such as the country where a specimen was found, may be inferred from the information on other specimens found during the same expedition. Currently, if one wants to retrieve all objects from the database that belong to an expedition, one would have to create a database query that contains the exact data boundaries of the expeditions and the names of all collectors involved. Either one of these bits of information is not enough, as the same group of biologists may have participated in an expedition more than once, and the database may also contain expeditions that overlap in time. In this paper a series of experiments is described to find a way to infer expedition information from the information available in the database. To this end, three approaches are compared: supervised machine learning, unsupervised machine learning, and rule-based methods.

[1] http://www.naturalis.nl

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 17–24,
Prague, 28 June 2007.

The obtained results vary, but prove that it is possible to extract the expedition information from the data at hand.

2 Related Work

The field of data mining, which is concerned with the extraction of implicit, previously unknown and potentially useful information from data (Frawley et al., 1992), is a branch of research that has become quite important recently as every day the world is flooded with larger amounts of information that are impossible to analyse manually. Data mining can, for instance, help banks identify suspicious transactions among the millions of transactions that are executed daily (Fayyad and Uthurusamy, 1996), or automatically classify protein sequences in genome databases (Mewes et al., 1999), or aid a company in creating better customer profiles to present customers with personalised ads and notifications (Linden et al., 2003). Knowledge discovery approaches often rely on machine learning techniques as these are particularly well suited to process large amounts of data to find similarities or dissimilarities between instances (Mitchell, 1997).

Traditionally, governments and companies have been interested in gaining more insight into their data by applying data mining techniques. Only recently , digitisation of data in the cultural heritage domain has taken off, which means that there has not been much work done on knowledge discovery in this domain. Databases in this domain are often created and maintained manually and are thus often significantly smaller than automatically generated databases from, for example, customers' purchase information in a large company.

This means it is not clear whether data mining techniques, aimed at analysing enormous amounts of data, will work for the data at hand. This is investigated here. Manual data typically also contains more spelling variations/errors and other inconsistencies than automatically generated databases, due to different persons entering data into the database. Therefore, before one can start the actual process of knowledge discovery, it is very important to carefully select, clean and model the data one wants to use in order to avoid using data that is too sparse (Chapman, 2003). This applies in particular to databases that contain large amounts of textual information, which are quite prevalent in the cultural heritage domain. Examples of textual databases can be found freely on the internet, such as the databases of the Global Biodiversity Information Facility[2], the University of St. Andrews Photographic Collection[3], and the Internet Movie Database[4].

3 Data

The data that has been used in this experiment is an animal specimen database from the Dutch National Museum for Natural History. The database currently contains 16,870 entries that each represent an object stored in the museum's reptiles and amphibians collection. The entries provide a variety of information about the objects in 37 columns, such as the scientific name of the object, how the specimen is kept (in alcohol, stuffed, pinned) and under which registration number, where it was found, by whom and under which circumstances, the name of the person who determined the species of the animal and the name of the person who first described the species. Most fields are rather compact; they only contain a numeric value or a textual value consisting of one or several words. The database also contains fields of which the entries consist of longer stretches of text, such as the 'special remarks' field, describing anything about the object that did not fit in the other database fields and 'biotope', describing the biotic and abiotic components of the habitat from which the object was collected. Dutch is the most frequent language in the database, followed by English. Also some Portuguese and German entries occur. Taxonomic values, i.e., the scientific names of the animal specimens, are in a restricted type of Latin. A snippet of the database can be found in Figure 1.

3.1 Data Construction

In order to be able the measure the performance of the approaches used in the experiments, the database was annotated manually with expedition information. Adding this information was possible because there was access to the original field books from which the database is made up. Annotating 8166

[2]http://www.gbif.org/

[3]http://special.st-andrews.ac.uk/saspecial/

[4]http://www.imdb.com/

Collector	Coll. Date	Coll. #	Class	Genus	Species	Country	Expedition
Buttikofer, J.	30-07-1881	424	Reptilia	Lamprolepis	lineatus	132	buttikoferliberia1881
Buttikofer, J. & Sala	09-10-1881	504	Amphibia	Bufo	regularis	132	buttikoferliberia1881
M. Dachsel	02-05-1971	971-MSH186	Reptilia	Blanus	mettetali	156	mshbrazil71
Hoogmoed, M.S.	04-05-1971	1971-MSH187	Reptilia	Quendenfeldtia	trachyblepharus	156	mshbrazil71
Hoogmoed, M.S.	09-05-1971	1971-MSH202	Reptilia	Lacerta	hispanica	156	mshbrazil71
C. Schuil	14-03-1972	1972-MSH35	Amphibia	Ptychadaena	sp.	92	mshghana72
P. Lavelle	-03-1972	1972-MSH40	Reptilia	Crotaphopeltis	hotamboeia	92	mshghana72
Hoogmoed, M.S.	23-03-1972	1972-MSH55	Amphibia	Phrynobatrachus	plicatus	92	mshghana72

Figure 1: Snippet of the animal specimen database

entries with this information took one person about 2 days. There were 8704 entries to which no expedition is assigned, either because these specimens were not collected during an expedition or because it was not possible to determine the expedition. These entries were excluded from the experiments . Expeditions which contained 10 or fewer entries were also excluded because these would make the data set too sparse. A total of 7831 database entries were used in this work, divided into 60 expeditions. Although the 'smallest' expeditions were excluded from the experiments, the sizes of the expeditions still vary greatly: between 2170 and 11 items (σ = 310.04). This is mainly due to the fact that new items are still added to the database continuously, in a rather random order, hence some expeditions are more completely represented than others.

The database contains several fields that contain information that will probably not be that useful for this work. Information that was excluded was the specimen's sex, the number of specimens (in cases where one database entry refers to several specimens, for instance kept together in a jar), how the animal is preserved, and fields that contain information not on the specimen itself or how it was found but on the database (e.g., when the database entry was added and by whom). Values from the 'altitude' and 'coordinates' fields were also not included in the experiments as this is information is too often missing in the database to be of any use (altitude information is missing in 85% of the entries and coordinates in 96%).

Some information in the database is repetitive; there is for instance a field called 'country' containing the name of the country in which a specimen was found, but there is also a field called 'country-id' in which the same information is encoded as a numerical value. The latter is more often filled than the 'country' field, which also contains values in different languages, and thus it makes more sense to only include values from the 'country-id' field. A small conversion is applied to rule out that an algorithm will interpret the intervals between the different values as a measure of geographical proximity between the values, as the country values are chosen alphabetically and do not encode geographical location.

In some cases it seemed useful to have an algorithm employ interval relations between numbers. The fields 'registration number' and 'collection number' were used as such. These fields sometimes contain some alphabetical values: certain collectors, for instance, included their initials in their series of collection registration numbers. These were converted to a numeric code to obtain completely numeric values with preservation of the collector information. This also goes for the fields in the database that contain information on dates, i.e., the 'date of determination', the 'date the specimen came into the museum' and the 'collection date' fields. The collection date is the most important date here as this directly links to an expedition. The other dates might provide indirect information, for instance if the collection date is missing (which is the case in 14%). To aid clustering, the dates were normalised to a number, possibly the algorithm benefits from the fact that a small numerical interval means that the dates are close together.

Person names from the 'author', 'collector', 'determiner', and 'donator' fields were normalised to a 'first name - last name' format. From values from the taxonomic fields ('class', 'order', 'family', 'genus', 'species', and 'sub species'), and 'town/village' and 'province/state' fields, as well as from the person name fields, capitals, umlauts, accents and any other non-alphanumerical characters were removed.

It proved that certain database fields were not suitable for inclusion in the experiments. This goes for the free text fields ‘biotope’, ‘location’ and ‘special remarks’. Treating these values as they are will result in data that is too sparse, as their values are extremely varied. Preliminary experiments to see if it was possible to select only certain parts of these fields did not yield any satisfying results and was therefore abandoned.

This resulted in feature vectors containing 18 features, plus the manually assigned expedition class.

4 Methodology

The majority of the experiments that were carried out in an attempt to infer the expedition information from the database involved machine learning. Therefore in this section three algorithms for supervised learning are described, followed by a clustering algorithm for unsupervised learning. This section is concluded with a description of the evaluation metrics for clusters used by the different approaches.

Algorithms

The first algorithm that was used is the ***k*-Nearest Neighbour** algorithm (*k*-NN) (Aha et al., 1991; Cover and Hart, 1967; DeVijver and Kittler, 1982). This algorithm is an example of a lazy learner: it does not model the training data it is given, but simply stores each instance of the training data in memory. During classification it compares the item it needs to classify to each item in its memory and assigns the majority class of the closest *k* (in these experiments *k*=1) instances to the new item. To determine which instances are closest, a variety of distance metrics can be applied. In this experiment the standard settings in the TiMBL implementation (Daelemans et al., 2004), developed at the ILK research group at Tilburg University, were used. The standard distance metric in the TiMLB implementation of *k*-NN is the Overlap metric, given in Equations1 and 2. Δ(X,Y) is the distance between instances X and Y, represented by *n* features, where δ is the distance between the features.

$$\Delta(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{1}$$

where:

$$\delta(x_i, y_i) = \begin{cases} abs & \text{if numeric, else} \\ 0 & \text{if} x_i = y_i \\ 1 & \text{if} x_i \neq y_i \end{cases} \tag{2}$$

The second algorithm that was used is the **C4.5** decision tree algorithm (Quinlan, 1986). In the learning phase, it creates a decision tree in a recursive top-down process in which the database is partitioned according to the feature that separates the classes best; each node in the tree represents one partition. Deeper nodes represent more class-homogeneous partitions. During classification, C4.5 traverses the tree in a deterministic top-down pass until it meets a class-homogeneous end node, or a non-ending node when a feature-value test is not represented in the tree.

Naive Bayes is the third algorithm that was used in the experiments. It computes the probability of a certain expedition, given the observed training data according to the formula given in Equation 3. In this formula υ_{NB} is the target expedition value, chosen from the maximally probably hypothesis ($\overset{argmax}{\upsilon_j \epsilon V} P(\upsilon_j)$, i.e., the expedition with the highest probability) given the product of the probabilities of the features ($\prod_i P(a_i|\upsilon_j)$).

$$\upsilon_{NB} = \overset{argmax}{\upsilon_j \epsilon V} P(\upsilon_j) \prod_i P(a_i|\upsilon_j) \tag{3}$$

For both the C4.5 algorithm and Naive Bayes the WEKA machine learning environment (Witten and Frank, 2005), that was developed at the University of Waikato, New Zealand, was used.

A quite different machine learning approach that was applied to try to identify expeditions in the reptiles and amphibians database is **clustering**. Clustering methods are unsupervised, i.e., they do not require annotated data, and in some cases not even the number of expeditions that are in the data. Items in the data set are grouped according to similarity. A maximum dissimilarity between the group members may be specified to steer the algorithm, but other than that it runs on its own. For an extensive overview of clustering methods see Jain et al., (1999). For this work, the options in choosing an implementation of a clustering algorithm were limited because many data mining tools are designed only for numerical data, therefore the WEKA machine learning environment was also used for the clustering experiments. As clustering is computationally expensive, it was only possible to run experiments with WEKA's implementation of the Expectation Maximisation (**EM**) algorithm (Dempster et al., 1977). Preliminary experiments with other algorithms indicated execution times in the order of months. The EM algorithm iteratively tries to converge to a maximum likelihood by first computing an expectation of the likelihood of a certain clustering, then maximising this likelihood by computing the maximum likelihood estimates of the features. Termination of the algorithm occurs when the predefined number of iterations has been carried out, or when the overall likelihood (the measure of how 'good' a clustering is) does not increase significantly with each iteration.

Cluster Evaluation

Since the data is annotated with expedition information it was possible to use external quality measures (Steinbach et al., 2000). Three different evaluation measures were used: **accuracy**, **entropy** (Shannon, 1948), and the **F-measure** (van Rijsbergen, 1979).

The evaluation of results for the supervised learning algorithms was calculated in a straightforward way: because the classifier knows which expeditions there are and which entries belong to which expedition, it checks the expeditions it assigned to the database entries to the manually assigned expeditions and reports the overlap as accuracy.

It gets a little bit more complicated with entropy. Entropy is a measure of informativity, i.e., the minimum number of bits of information needed to encode the classification of each instance. If the expedition clusters are uniform, i.e., all items in the cluster are very similar, the entropy will be low. The main problem with using entropy for evaluation of clusters is that the best score (an entropy of 0) is reached when every cluster contains exactly on instance. Entropy is calculated as follows: first, the main class distribution, i.e., per cluster the probability that a member of that cluster belongs to a certain cluster, is computed. Using that distribution the entropy of each cluster is calculated via the formula in Equation 4. For a set of clusters the total entropy is then computed via the formula in Equation 5, in which *m* is the total number of clusters, s_y the size of cluster *y* and *n* the total number of instances.

$$E_y = -\sum_x P_{xy} log(P_{xy}) \tag{4}$$

$$E_{total} = \sum_{y-1}^{m} \frac{s_y \cdot E_y}{n} \tag{5}$$

The F-measure is the harmonic mean of precision and recall, and is commonly used in information retrieval. In information retrieval recall is the proportion of relevant documents retrieved out of the total set of relevant documents. When applied to clustering a 'relevant document' is an instance that is assigned correctly to a certain expedition, the set of all relevant documents is the set of all instances beloning to that expedition. Precision is the number of relevant documents retrieved from the total number of documents. So when applied to cluster evaluation this means the number of instances of an expedition that were retrieved from the total number of instances (Larsen and Aone, 1999). This boils down to Equations 6 and 7 in which *x* stands for expedition, y for cluster, n_{xy} for the number of instances belonging to expedition *x* that were assigned to cluster *y*, and n_x is the number of items in expedition *x*.

$$Recall(x, y) = \frac{n_{xy}}{n_x} \tag{6}$$

$$Precision(x, y) = \frac{n_{xy}}{n_y} \tag{7}$$

The F-measure for a cluster *y* with respect to expedition *x* is then computed via Equation 8. The F-measure of the entire set of clusters is computed through the function in Equation 9, which takes the weighted average of the maximum F-measure per expedition.

$$F(x, y) = \frac{2 \cdot Recall(x, y) \cdot Precision(x, y)}{Precision(x, y) + Recall(x, y)} \tag{8}$$

$$F = \sum_x \frac{n_x}{n} \max\{F(x, y)\} \tag{9}$$

5 Experiments and Results

First, two baselines were set to illustrate the situation if no machine learning or other techniques would be applied to the database. if one were to randomly assign one of the 60 expeditions t the entries this would go well in 1.7% of the cases. If all entries were labelled as belonging to the largest expedition this would yield an accuracy of 28%. In all machine learning experiments 10-fold cross validation was used for testing performance.

A series of supervised machine learning experiments was carried out first to investigate whether it is possible to extract the expeditions during which the animal specimens were found at all. Three learning algorithms were applied to the complete data set, which yielded accuracies between 88% and 98%. Feature selection experiments with the C4.5 decision tree algorithm indicated that features 'town/village', 'collection number', 'registration number', 'collector' and 'collection date' were considered most informative for this task, hence the experiments were repeated with a data set containing only those features. The results of both series of experiments are to be found in Table 1. For the C4.5 and Naive Bayes experiments the accuracy deteriorates significantly when using only the selected features ($\alpha = 0.05$, computing using McNemar's test (McNemar, 1962)), but it stays stable for the *k*-NN classifier. This indicates that not all data is needed to infer the expeditions, but that it matters greatly which approach is taken. However, as neither of the algorithm benefits from it, feature selection was not further explored.

Algorithm	All feat.	Sel. feat.
k-NN	95.9%	95.9%
C.4.5	98.3%	94.4%
NaiveBayes	88.1%	73.5%

Table 1: Accuracy of supervised machine learning experiments using all features and selected features

In these experiments all database entries were annotated with expedition information, which in a real setting is of course not the case. Through running a series of experiments with significantly smaller amounts of training data it was found that by using only as little as 5% of the training data (amounting to 392 instances) already an accuracy of 85% is reached. Annotating this amount of data with expedition information would take on person less than an hour. By only using 45% of the training data an accuracy of 97% is reached[5]. In Figure 2 the complete learning curve of the *k*-NN classifier is shown.

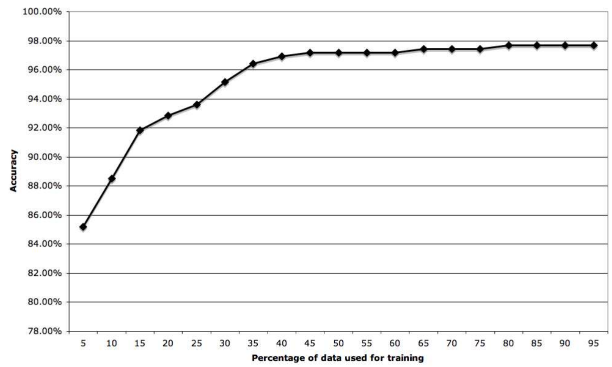

Figure 2: Accuracy of *k*-NN per percentage of training data

Ideally, one does not want to annotate data at all, therefore the use of a clustering algorithm was explored. For this, the EM algorithm from the WEKA machine learning environment was used. The result, as shown in Table 2, is not quite satisfying, but still well above the set baselines. As can be seen in Table 2, the clustering algorithm does not come up with anywhere near as many clusters as needed and unfortunately WEKA does not present the user with many options to remedy this. An intermediate experiment between completely supervised and unsupervised machine learning was attempted, i.e., prespecifying a number of clusters for the algorithm to define, but this was computationally too expensive to carry out.

Algorithm	# Clusters	Accuracy
EM	7	46.0%

Table 2: Result of clustering experiment

Since the clustering algorithm does not achieve an accuracy that is satisfying enough to use in a real setting and supervised learning requires annotated data, also a traditional, and quite different approach

[5]The slightly higher achieved accuracy in the learning curve experiments is due to the fact that the learning curve was not computed via cross-validation

was tried: namely finding expeditions via rules. Via a couple of simple rules the data set was split into possible expeditions using only information on collection dates, collector information and country information.

1. Sort dates in ascending order, start a new expedition when the distance between two sequential dates is greater than the average distance of the collection dates

2. First, sort collector information in ascending order, then sort collection dates in ascending order, start a new expeditions when the distance between two dates is greater than the average distance between dates or when a new collector is encountered

3. First, sort by country information, then by collector, and finally by collection date, start a new expedition when country or collectors change, or when the distance between two dates is greater than the average distance between dates

Surprisingly, only grouping collection dates already yields an F-measure of .83. This includes 1299 entries that contain no information on the collection date, leaving those out would increase precision on the entries whose collection date is not missing to an F-measure of .94. In Table 3 results of the rule-based experiments are shown. It is expected that when the database is further populated the date-rule will work less well as there will be more expeditions that overlap. The date+collector-rule should remedy this, although it does not work very well yet as spelling variations in the collector names are not taken into account at the moment.

Rules	# Clusters	F-measure	Entropy
1	78	.83	.16
2	199	.75	.15
3	216	.73	.11

Table 3: Results of the rule-based experiments

6 Conclusions and Future Work

In this work various approaches were presented to rediscover expedition information from an animal specimen database. As expected, the supervised learning algorithms performed best, but the disadvantage in using such an approach is the requirement to provide annotated data. However, a series of experiments to gain more insight into the quantities of data necessary for a supervised approach to perform well, indicate that only a small set of annotated data is required in this case to obtain very reasonable results. If no training data is available, a rule-based approach is a realistic alternative. Although it must be kept in mind that rules need to be created manually for every new data set. For this data set relatively simple rules already proved to be quite effective, but for other data sets deriving rules can be much more complicated and thus more expensive. This particular set of rules is also expected to behave differently when the database is extended with more entries from overlapping expeditions.

For the experiments presented in this work, only entries from the database of which the expedition they belonged to was known were used, which constitutes only half of the database entries. Researchers at Naturalis estimate that about 30% of the database entries do not belong to an expedition, while the other 20% not included here belong to unknown expeditions. The decision to exclude the expedition-less entries was made as these entries would imbalance the data and impair evaluation as it would not be possible to check predictions against a ‘real value’. If all database entries would belong to a known expedition the performance of the approaches described in this paper that satisfactory results could be achieved over the complete data set. To prove this hypothesis one would need to test the approaches on other data sets which can be completely annotated. Performing such tests might provide more insight into how well the approaches would deal with a data set where all entries have an associated expedition. The natural history museum has several other similar (but smaller) data sets, which might be suitable for this task, and which will be tested as part of future work for evaluating the approaches described here. It may also be interesting to investigate what can be inferred from the other fields defined in other data sets.

A less satisfying aspect of the research described in this paper is that many of the intended experiments with unsupervised machine learning were too computationally expensive to be executed. Potential workarounds to the limitation of certain implementations of clustering algorithms, in that they only work on numeric data, are sought in converting the textual data to numeric values and in the investigations into implementations of algorithms that can deal with textual data.

A particular peculiarity of textual data, from which the rule-based approach suffers, is the fact that the same name or meaning can be conveyed in several ways. Spelling variations and errors were for instance not normalised. Hence the approaches treated 'Hoogmoed' and 'M S Hoogmoed' as two different values whereas they may very well refer to the same entity.

From this work it can be concluded that the expedition information can definitely be reconstructed from the animal specimen database that was used here, but for it to be used in a real world application it still needs to be tested and fine-tuned on other data sets and extended to be able to deal with entries that are not associated with any expedition.

Acknowledgments

The research reported in this paper was funded by NWO, the Netherlands Organisation of Scientific Research as part of the CATCH programme. The author would like to thank the anonymous reviewers, and Antal van den Bosch and Caroline Sporleder for their helpful suggestions and comments.

References

David W. Aha, Dennis Kibler, and Mark K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Arthur D. Chapman. 2003. Notes on Environmental Data Quality-b. Data Cleaning Tools. Internal report, Centro de Referência em Informação Ambiental (CRIA).

T. M. Cover and P. E. Hart. 1967. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13:21–27.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2004. Timbl: Tilburg memory based learner, version 5.1, reference guide. Technical Report 04-02, ILK/Tilburg University.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodology)*, 39(1):1–38.

P. A. DeVijver and J. Kittler. 1982. *Pattern recognition. A statistical approach*. Prentice-Hall, London.

U. Fayyad and R. Uthurusamy. 1996. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26.

William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. 1992. Knowledge discovery in databases: An overview. *AI Magazine*, 13:57–70.

A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September.

Bjorner Larsen and Chinatsu Aone. 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of KDD-99*, San Diego, CA.

G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan/Feb.

Q. McNemar. 1962. *Psychological Statistics*. Wiley, New York.

H. W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. 1999. Mips: a database for genomes and protein sequences. *Nucleic Acids Research*, 27(1):44–48.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. Technical report, Department of Computer Science, University of Minnesota.

Cornelis Joost van Rijsbergen. 1979. *Information Retrieval*. Buttersworth.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.