Frequency-Based vs. Knowledge-Based Similarity Measures for Categorical Data

Summaya Mumtaz and Martin Giese
Department of Informatics,
University of Oslo, Norway
{summayam@ifi.uio.no, martingi@ifi.uio.no}

Abstract

Calculation of similarity between two entities is a key step in several data mining processes. While there are several common similarity measures for continuous data, there is little work for categorical data. Most approaches are purely data-driven and don't consider the inherent dependencies of complex domains such as geological structures, phylogenetics, etc. We propose two new similarity measures that take into account semantic information to calculate the similarity between two categorical values. Semantic information is represented as a hierarchy extracted from an ontology or a domain taxonomy. The first approach calculates semantic similarity by combining the data-driven approach with the hierarchy imposed on the possible categorical values. The second approach ignores the data and uses only the hierarchy to calculate semantic similarity. We apply our methods to a specific complex data mining task in the oil and gas industry: reservoir analogue identification. The two proposed measures are compared to existing data-based measures.

1 Introduction

The context of this work is the combination of data-based (statistical) methods with knowledge-based methods in data science. In many disciplines, there is a considerable body of domain knowledge available, while data sets may not always be large enough to support machine learning of complex relationships. In this work, we look specifically at *similarity measures* (or equivalently distance measures), which lie at the core of a number of machine learning tasks such as clustering, outlier identification and classification (k-NN). We concentrate on entities described by *categorical* data, feature values taken from a finite set of possible values with no inherent order. The domain knowledge we wish to incorporate is given in the form of hierarchies that can be extracted from domain ontologies, standard classifications, etc.

There is a variety of suitable metrics to quantify similarity for numerical data such as Euclidean or Manhattan distance (Esposito et al. 2000). These methods are not directly applicable to non-numerical data. However, defining sensible metrics for categorical attributes is challenging.

The most common approach in machine learning algorithms for handling categorical data is one-hot encoding (Alkharusi 2012; Davis 2010). A binary column is created for each unique value of the categorical column. This yields a high-dimensional sparse matrix, containing a significant proportion of zeros. This approach requires high computational resources, is unable to handle unseen values and ignores any domain dependencies known to exist between values of the same categorical attribute.

In a supervised learning approach, the distance $\delta(x, y)$ between two categorical values can be defined using value distance matrix (Stanfill and L. Waltz 1986) and modified value distance matrix (Cost and Salzberg 1998).

For unsupervised learning, the hamming distance is used and similarity is defined as a matching measure that assigns 1 if both values are identical, and 0 otherwise (Esposito et al. 2000; Ahmad and Dey 2007). Various similarity measures have been derived using this distance measure, e.g. Jaccard similarity coefficient, Sokal-Michener similarity measure, Grower-Legendre similarity measure, etc. (Esposito et al. 2000). These measures are inherently quite coarse: in the absence of an ordering between the categorical values, the only possible distinction is whether two values are identical or not (Esposito et al. 2000).

To improve on these, *frequency-based* similarity measures have been proposed that take the frequency distribution of different attribute values into account. These measures are data-driven and hence are dependent on certain data characteristics such as the size of data, number of attributes, number of values for each attribute and distribution of frequency of each value. While data-driven measures perform well on simple datasets, these measures are unable to take into account semantic relationships and often don't perform well on complex datasets with hidden domain dependencies. Moreover, a concept of similarity that is based solely on how often values occur in the data cannot be expected to give reasonable results in all cases. Using frequencies seems more like a 'last straw' when frequencies are the only distinguishing feature between categorical values.

In this paper, we propose an alternative way to measure similarity for categorical data in an unsupervised setting. We combine a frequency-based measure with explicitly represented domain knowledge given in the form of a hierarchy on attribute values, and we also consider a measure that is based purely on the hierarchy, without taking frequencies into account.

Section 2 describes the related work. Section 3 explains the problem formulation and proposed algorithm. Section 4 presents the dataset and evaluation by comparing with existing algorithms.

2 Literature Review

The surveys (Boriah, Chandola, and Kumar 2008; Alamuri, Surampudi, and Negi 2014) discuss various similarity measures for categorical data. Wilson and Martinez (Wilson and Martinez 2000) have studied in-depth heterogeneous functions for mixed data (categorical and continuous variables) for instance-based learning. Their approach is based on supervisor learning where each instance has class labels in addition to input variables. The focus of this paper is to find similarity in an unsupervised setting where information regarding classes is unknown.

For unsupervised learning, various techniques have been proposed (Boriah, Chandola, and Kumar 2008). The majority of these techniques are based only on the data-driven approach. However, in some other domains like in natural language processing, research is being conducted to calculate similarity based on semantics and domain knowledge. Below, we provide an overview of the existing data-driven measures, followed by research done in natural language processing.

The simplest similarity measure used is known as overlap measure (Boriah, Chandola, and Kumar 2008). Similarity of 1 is assigned when two categorical values are identical otherwise similarity is assigned as 0. The overall similarity between two data instances of multivariate categorical data is proportional to the number of attributes in which they are identical. The overlap measure does not distinguish different values of attributes hence matches and mismatches are treated equally. Goodall proposed a similarity measure to normalize similarity between two data instances by the probability of occurrences in a random sample (W. Goodall 1966). This measure assigns a higher similarity score to the values which are less frequent. Gambaryan proposed similarity measure by giving more weight to matches where the frequency of occurrence of categorical values is about half in the dataset (Gambaryan 1964). (Eskin et al. 2002) developed a normalization kernel intrusion detection system. This measure assigns more weight to mismatches of attributes that contain many values. Inverse Occurrence frequency (IOF) assigns lower similarity values to mismatches that are based on more frequent values. IOF measure is derived from information retrieval (Sparck Jones 2004) and is associated with the idea of inverse document frequency. The Occurrence frequency (OF) measure assigns lower similarity to mismatches on less frequent values and mismatches on more frequent items are assigned higher similarity (Boriah, Chandola, and Kumar 2008).

Lin proposed a similarity framework based on information theory (Lin 1998). According to Lin, similarity can be explained in terms of a set of assumptions. If the assumptions are considered true, the similarity measure is necessarily followed. Therefore, the similarity between the two values is calculated by the ratio between the amount of information required to state the commonality of both values and the information needed to fully describe both values separately. Lin derived similarity measure for words, ordinal and string data.

Das and Mannila's research is based on the key point that attribute value similarity is related to other attributes (Das and Mannila 2000). They proposed Iterated Contextual Distances (ICD) based on the idea that attribute and object similarities are interdependent. ICD finds attribute similarity, sub relation, and row similarity. Ahmed and Dey proposed a distance-based measure in term of co-occurrence of values, the overall distribution of two attribute values are considered along with their co-occurrence with the values of other attributes (Ahmad and Dey 2007).

Document or sentence similarity is considered the basic task for many natural language processing(NLP) engines such as information retrieval, query answering, and text summarization. Semantic-based methods use information from dictionaries (WordNet) to find relatedness between two terms. Classic methods in NLP are based on the shortest path measure (Roy et al. 1989). (Leacock and Chodorow 1998) proposed a similarity technique based on the shortest path between nodes in a taxonomy and the number of nodes.(Huang and Sheng 2012) based their sentence similarity measure by using WordNet information content and string edit distance, for paraphrase recognition.

However, the techniques mentioned above are not directly suitable for categorical features. In an NLP setting, there are many terms in a complete sentence or document, that provide the neighborhood context and aid understanding the semantics. Furthermore, NLP tasks are constrained by the sentence structures and grammar of a particular language such as the ordering of subject, verb, noun, etc. However, categorical features are represented by single domain terms with no obvious representation of neighborhood or the context that explains the semantic similarity. The main focus here is to define semantic similarity between categorical terms based on the characteristics extracted from domain knowledge.

3 Problem Formulation

In this section, we first discuss a toy example to identify the drawbacks of frequency-based similarity approaches. Further, we provide an overview of metric properties and semantic similarity to establish the foundation of the proposed similarity measure.

We analyze the problems in existing work and inherent challenges associated with categorical data based on the toy dataset in Table 1. The dataset consists of candidates' profiles and we wish to retrieve matching candidates for a given job advertisement.

Many of the data-driven similarity measures consider two values of a given categorical attribute to be similar if both have similar frequency distributions. For instance, the OF similarity measure for values of an attribute is calculated as follows (Boriah, Chandola, and Kumar 2008).

Table 1: Toy Dataset

User ID	Occupation	Education
1	Computer Programmer	Bachelors
2	Administrative Staff	Bachelors
3	HR Manager	Bachelors
4	HR Manager	Masters
5	Software Developer	Bachelors
6	Computer Programmer	Masters

$$OF(x,y) = \begin{cases} 1 & \text{if } x = y \\ \frac{1}{(1+\log(\frac{N}{f(x)+1})+\log(\frac{N}{f(y)+1}))} & \text{if } x \neq y \end{cases} \tag{1}$$

where $f(x)$ is the number of occurrences of the attribute value x and N represents the total number of observations in the data set. Similarity between pairs 'Computer Programmer' and 'HR Manager' and 'Computer Programmer' and 'Software Developer' based on equation 1 is calculated as:
$OF(\text{Comp. Programmer}, \text{HR Manager}) = 0.64$
$OF(\text{Comp. Programmer}, \text{Soft. Developer}) = 0.44$

These numbers would indicate that the Programmer is more similar to HR Managers than to Developers. However, based on the evaluation of semantic evidence observed in a knowledge source (such as an ontology or a standard classification) shown in Table 2, it is evident that computer programmers and software developers perform the same work activities and tasks hence having a greater semantic similarity.

Semantic similarity can be made explicit in different ways, and one of the prominent ways is through hierarchies, which we will use in this paper. Section 3.1 explains in detail the formal definition of hierarchies.

3.1 Hierarchies

Our similarity measures are based on a given hierarchical structure of the value range of categorical features. Formally, we assume that the categorical values for each feature form a finite, partially ordered set (poset). A poset is an ordered pair of binary relation $\sqsubseteq$ defined over a set S, such that $(\sqsubseteq, S)$ satisfies the following properties: Let $x, y, z \in S$,

- Reflexivity: $x \sqsubseteq x$
- Antisymmetry: if $x \sqsubseteq y$ and $y \sqsubseteq x$, then $x = y$
- Transitivity: if $x \sqsubseteq y$ and $y \sqsubseteq z$, then $x \sqsubseteq z$

If $a \sqsubseteq b$, we call b an ancestor of a. The intention of $a \sqsubseteq b$ is that b is in some way more general, broader, etc. than a. E.g., for the occupations in Fig. 1, TopExecutives $\sqsubseteq$ ManagementOccupations; for data about geographic areas, we could have Oslo $\sqsubseteq$ Norway $\sqsubseteq$ Europe.

If domain knowledge is given in the form of an ontology, in some cases (depending on the modeling style), the relation $\sqsubseteq$ will correspond to parts of the *is-a* subclass relation of the ontology, but in others it won't. E.g. it doesn't make sense to consider Norway a sub-class or sub-concept of Europe, but it still makes sense to consider a hierarchy of geographic regions.

A value $c \in S$ is called a *lowest common ancestor* of two node values $a \in S$ and $b \in S$ if $c \in S$ is the lowest (i.e. deepest) node that has both $a \in S$ and $b \in S$ as descendants. It is the first shared ancestor of a and b located farthest from the root. In a hierarchy two values have a lowest common ancestor denoted as $a \sqcup b$. A value is called a *leaf value* if it is not the ancestor of any other value.

In this paper, we add a restriction to our hierarchies by only considering mono-hierarchies: we assume that there is some root value r in the hierarchy, such that $a \sqsubseteq r$ for all $a \in S$, and that all values except the root have exactly one direct ancestor. In other words, the hierarchy is tree-shaped.

3.2 Semantic Similarity

Semantic similarity refers to similarity based on meaning or semantic content as opposed to form (Smelser and Baltes 2001). Semantic similarity measures are automated methods for assigning a pair of concepts a measure of similarity and can be derived from a taxonomy of concepts arranged in is-a relationships (Pedersen, Pakhomov, and Patwardhan 2005). The concept of semantic similarity has been applied in Natural language processing for the past decade to solve tasks such as the resolution of ambiguities between terms, document categorization or clustering, word spelling correction, automatic language translation, ontology learning or information retrieval. Similarity computation for categorical data can improve the performance of existing machine learning algorithms (Ahmad and Dey 2007) and may ease the integration of heterogeneous data (Wilson and Martinez 2000).

Is-a relationships in a concept hierarchy encompass formal classification, properties and relations between concepts and data. This provides us with a common understanding of the structure of a domain, explicit domain assumptions and reuse of domain knowledge. In order to achieve interpretable and good quality results in machine learning models, it is vital to take this information into account. This intuition motivates us to link the notion of similarity based on is-a relationships with the similarity measures for categorical data. We develop a framework to use is-a relationships extracted from a concept hierarchy to quantify semantic similarity and propose a distance measure for categorical data.

3.3 Proposed Framework

In this paper, we propose two techniques for measuring similarity based on domain knowledge, extracted as the concept hierarchy. First, we present a framework for calculating semantic similarity using information content and concept hierarchy by modifying Resnik's idea (Resnik 1970). To compare the performance of information-content based semantic measure, we extended the idea to introduce a simple similarity measure based only on concept hierarchy.

Further, we are interested in computing global semantic similarity in a multi-dimensional setting where we have several hierarchy-structured features. We define the global similarity between two data objects X and Y in a d-dimensional

Table 2: Occupation Activities and Skills

Occupation	Work Activity	Skills
HR Manager	Liaise between departments	PeopleSoft, SAP
Computer Programmer	Write programming code	C++, Java, Python
Software Developers	Modify software programs	C++, Oracle ,Python

attribute space as,

$$\delta(X, Y) = \sum_{i}^{d} w_i \delta(x_i, y_i) \tag{2}$$

where $\delta(x_i, y_i)$ corresponds to similarity between two values x and y in the i-th dimension and w_i is the weight associated with each dimension. The following section presents both frameworks for calculating semantic based similarity $\delta(x_i, y_i)$.

Information Content Semantic Similarity (ICS) This approach is based on a modification of Resnik's idea (Resnik 1970). Resnik proposed a measure for finding semantic similarity in an is-a taxonomy based on information content and defined similarity between two nodes in a hierarchy as the extent to which they share common information.

In order to formulate the semantic similarity of two given categorical values, the key intuition is to find the common information in both values. This information is represented by the lowest common ancestor in the hierarchy that subsumes both values (Lin 1998). If the lowest common ancestor of two values is close to leaf nodes, that implies both values share many characteristics. As the lowest common ancestor moves up in the hierarchy, fewer commonalities exist between a given pair of values.

For the given dataset, we can map the 'Occupation' attribute to the O*net taxonomy[1](Fig. 1) by placing all the values at the corresponding leaf nodes in the occupation hierarchy whereas intermediate nodes represent the lowest common ancestors for given pairs. In Fig. 1[2],'Computer Programmer' and 'Software Developer' are both subsumed by the lowest common ancestor 'Computer Occupations', whereas the lowest common ancestor that subsumes the concept 'HR Manager' and 'Computer Programmer' is 'Occupation'(root node of the occupation hierarchy). Hence, taking into account the lowest common ancestor, we expect that the similarity between Computer Programmer and Software Developer to be significantly greater than the similarity between the Computer Programmer and HR Developer.

Our intuition about the concept of semantic similarity is that for two categorical values x and y that share lowest common ancestor c, farthest from the root node, are always considered to be more semantically similar than to two categorical values x and z that share lowest common ancestor c' close to root node. In addition, identical values should have a maximum similarity of 1.

In order to formulate the semantic similarity of values based on the lowest common ancestor, we use the idea of associating probabilities with the values (Resnik 1970). We base ourselves on a function $p : C \to [0, 1]$ such that for any $c \in S$, $p(c)$ represents the probability of the feature value being $\sqsubseteq c$. Furthermore, using information theory we can state that the information content of a feature having some value is quantified as negative of log likelihood (Ross 1976).

For categorical data, we can find the information content I of the lowest common ancestor c by finding the information content of all the leaf values subsumed by c in the hierarchy.

$$I(c) = -log \sum_{n \in \ leaf(c)} p(n) \tag{3}$$

where $leaf(c)$ is the set of all leaf values in $x \in S$ such that $x \sqsubseteq c$. The probability of leaf values may be estimated by the relative frequency.[3]

$$p(n) = \frac{frequency(n)}{N} \tag{4}$$

where N is the number of samples.

Based on the above definitions, we formulate information content based semantic similarity(ICS) between two categorical values x and y as

$$Sim(x, y) = \begin{cases} 1 & \text{if } x = y. \\ \frac{I(x \sqcup y)}{\max(I(x \sqcup y))} & \text{else if } x \neq y \end{cases} \tag{5}$$

where $I(x \sqcup y)$ denotes the information content of the lowest common ancestor of x and y, calculated by using equation 3 and $\max(I(x \sqcup y)$ represents the maximum information content of all given pair of leaves and is used for normalization.

Hierarchy-based Semantic Similarity(HS) As explained earlier, the main intuition of semantic similarity is based on the idea that any two values having the lowest common ancestor close to leaf nodes, should have high similarity and vice versa. Hence, we quantify semantic similarity by considering the level of the lowest common ancestor in the hierarchy. The level of a node is defined by $1+$ the number of connections between the node and the root[4]. Greater the level of the lowest common ancestor of any given pair of values in the hierarchy, more similar the values are. We formulate the similarity as,

$$Sim(x, y) = \begin{cases} 1 & \text{if } x = y. \\ \lambda^{d - level(x \sqcup y)} & \text{else if } x \neq y \end{cases} \tag{6}$$

[1] https://www.onetcenter.org/taxonomy.html

[2] https://www.bls.gov/soc/soc_structure_2010.pdf

[3] Probabilities may also be known from other sources, for instance known priors for the specific domain.

[4] Level starts from 1 and the level of the root is 1

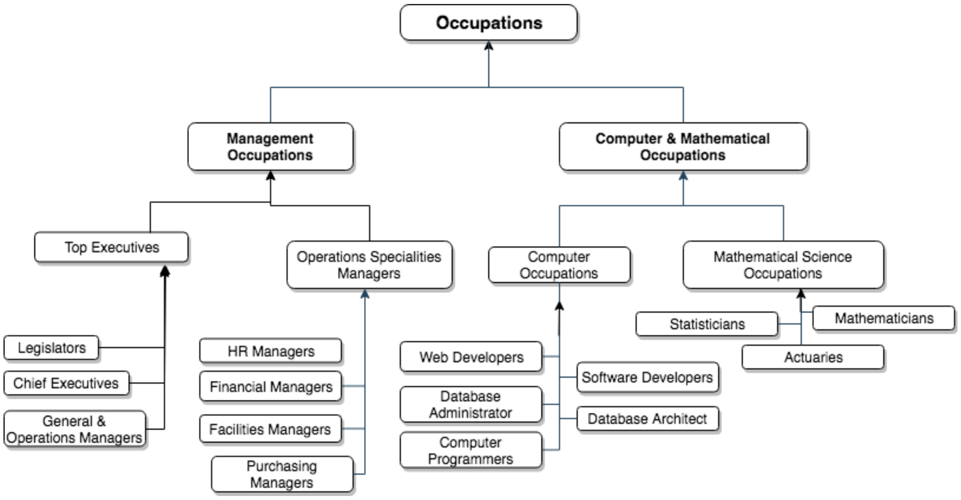

Figure 1: O*net Occupation Taxonomy

Where $0 < \lambda < 1$ is a fixed decay parameter, $level(n)$ is the distance of n from the root in the hierarchy, and $d = \max_{n \in X} level(n)$ is the maximum depth of the hierarchy.

The main advantage of equation 6 is that the calculation of semantic similarity no longer requires any input from training data such as information content. Once the concept hierarchy is formalized, we can measure the similarity between any two values including the categorical values not observed in the training data.

Below, we explain how to perform evaluation of the proposed techniques.

4 Evaluation

In this section, we compare the ICSD and HDM approaches to other similarity measures for the identification of reservoir analogues of a target reservoir, given a dataset of known reservoirs. This use-case is further explained in Section 4.2 below.

4.1 Baseline Methods

The following four state-of-the-art similarity/distance measures are compared with the proposed techniques: Occurrence Frequency (OF) (Boriah, Chandola, and Kumar 2008), Eskin Similarity measure (Boriah, Chandola, and Kumar 2008; Eskin et al. 2002) , Lin Similarity measure (Lin 1998) and Coupled Similarity Matrix (CMS) (Jian et al. 2018).

We compare the performance of the different similarity measures in a recommendation scenario: given a query item, we compute its similarity to each item in the 'training' dataset using Equation 2, and determine the top k items with highest similarity.

For our evaluation, we do this for all of the different similarity measures, and compare the outcome to a fixed 'gold standard' list of items to determine the average precision.

For our experimental evaluation, we have chosen reservoir analogues (explained in the section below): a complex task in the Oil and Gas industry. To the best of our knowledge, there exists no standard machine learning system for solving this use case. The common industrial practice to date is to conduct a manual analysis by human experts.

4.2 Reservoir Analogues

In the Oil and Gas Industry, during the exploration phase, analogous reservoirs are used to study reservoirs that lack critical information. Any reservoir with a deficit of critical information is known as a "target reservoir", and "analogous reservoirs" are ones expected to have similar characteristics. (Martín Rodríguez et al. 2013).

Usually, a technical evaluation team must analyze various data types – seismic, well logs, test, and cores – in order to make the first approximation of analogous reservoirs. Due to a lack of resources and time constraints, the first approximation is usually the neighboring reservoirs to provide an estimate of the fluid and rock properties of the target reservoir. A single analogue is mostly used because it is in the same geographic region or basin. This is risky however, since it does not always give sufficient information to characterize a new prospect. Furthermore, it becomes a tedious task for new target reservoirs where no neighboring reservoir exists.

Limited efforts have been made to identify analogues based on machine learning (Martín Rodríguez et al. 2013; Perez-Valiente et al. 2014). In order to generate a list of ranked reservoirs based on similarity, it is important to automate this process using a standard knowledge source and to develop a method that is flexible enough to produce analogues for reservoirs with no neighboring analogues.

4.3 Dataset

The main source of information used in this evaluation is the dataset of reservoirs licensed by IHS[5]. It comprises a total of 43000 reservoirs and various properties/attributes associated with each reservoir. According to domain experts, only a few key parameters are known during the initial stage of

[5] https://ihsmarkit.com/index.html

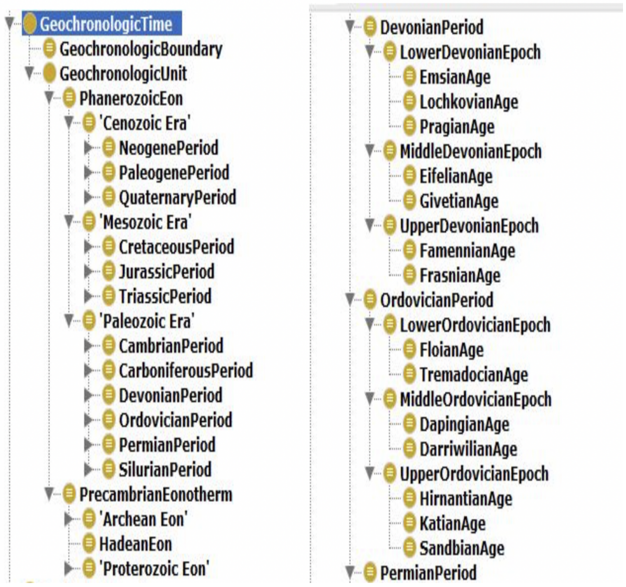

Figure 2: The hierarchy of geologic age.

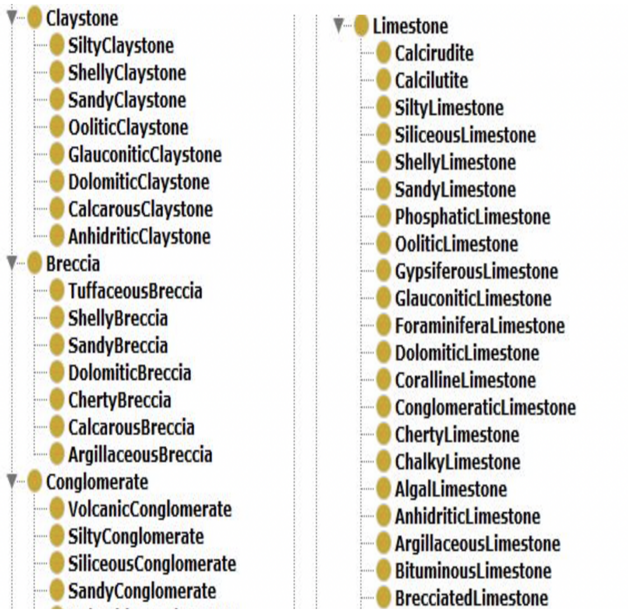

Figure 3: Ontology showing IS-A relationships for Lithology

reservoir identification. Hence for our analysis of retrieving similar reservoirs, we use the following set of key parameters/attributes identified by domain experts.

- Depositional Environment
- Lithology
- Age
- Geographical Location
- Structural Setting

Detailed definitions of these parameters are described in the section below.

4.4 Semantic Information for Attributes

This section explains the process of standardizing the semantic information used in the calculation of similarity. Due to data confidentiality, we only explain two attributes 'Age' and 'Lithology.'

Reservoir Age: A geologic age is a subdivision of geologic time that divides an epoch into smaller parts. A succession of rock strata laid down in a single age on the geologic timescale is a stage. The geological time has been divided into eras, periods and epochs. The named divisions of the geological time are based on fossil evidence. Fig. 2 shows a part of an ontology developed to show how geological times are organized into Erathem, Period, Epoch and Age.

Note that age can also be given on a linear scale, e.g. in millions of years. However, the characteristics of rocks deposited in different geologic eras, periods, and epochs differ so much that their position in the hierarchy is a much better indicator of similarity than the numerical difference in age.

Lithology: The lithology of a rock unit is a description of its physical characteristics visible at outcrop, in hand or core samples or with low magnification microscopies, such as color, texture, grain size, or composition. There is no standard ontology for lithology. With the help of geologists, we develop an ontology that considers all the categorical values occurring in data and groups them based on similar physical characteristics. In Fig. 3, we show a part of this ontology.

4.5 Data Pre-processing

The main challenge associated with the given data is a large number of categorical values associated with each attribute. For the attribute 'Age,' there are about 250 unique values. These values are not standardized. Hence, there are instances where the same category exists in the dataset with various names. Furthermore, most of the age values are unofficial names, which are used only in a few specific areas of the world. With the help of geological experts, we replaced these unofficial names by standard domain names.

For the attribute 'Depositional Environment,' there are 32 unique values occurring in the given data set. Some categorical values are merged together based on the same geological properties identified by domain experts.

In the original data set, there are 1731 categories of the attribute 'Lithology.' The raw values of lithology contain abbreviations for the same lithology, unofficial lithology names, and combinations of various lithologies. These categories are replaced with the standard names and combinations are replaced with only primary lithology, which leads to 228 unique categories.

Outliers are extreme values that deviate from other observations on data, they may indicate variability in measurement, experimental errors or a novelty. In order to avoid the disastrous effect on the results of the statistical analysis, a step is added to identify, analyze and delete outliers in the dataset. In this step, for every attribute, we remove the values that don't confirm with standard domain names.

After cleaning the data, the comparative evaluation between ICS, HS and existing similarity algorithms is conducted.

4.6 Evaluation Method

For the given task, we will evaluate the similarity measure on two main objectives.

- Retrieving top 15 similar analogues to the target reservoir.
- Producing the result in a ranked order such that the first retrieved analogue corresponds to the most similar reservoir to the target reservoir.

Mean Average Precision (MAP) is the mostly commonly used evaluation metric in information retrieval and object detection (Baeza-Yates and Ribeiro-Neto 2008). MAP is the arithmetic mean of the average precision (AP) values for an information retrieval system over a set of n query topics (Liu Ling 2009) . It can be expressed as follows:

$$MAP = \frac{1}{n} \sum_{n} AP_n \tag{7}$$

Precision for a classification task is defined as

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{8}$$

Based on Equation 8, recommender system Precision (P) is defined as,

$$P = \frac{\textit{\# of our recommendations that are relevant}}{\textit{\# of items we recommended}} \tag{9}$$

For evaluating the performance of recommender systems, we are only interested in recommending top-N items to the user. Usually, the higher the number of relevant recommendations at the top, the more positive is the impression of the users. Therefore, it is sensible to compute precision and recall metrics in the first N items instead of all the items. Thus the precision at a cutoff k is introduced in order to evaluate ranking, where k is an integer that is set by the user to match the objective of the top-N recommendations. Average precision at cutoff k, is the average of all precisions in the places where a recommendation is a true positive and is defined as follows:

$$AP_q@K = \frac{1}{K} \sum_{i=1}^{K} P(i) \cdot Rel(i) \tag{10}$$

where K represents the top K recommendations for the given query q and $Rel(i)$ shows the relevance of the recommendation. $Rel(i)$ is 1 if the recommended item was relevant(true positive) otherwise 0.

Table 3: Average Precision for the selected target reservoirs

Reservoir	**ICS**	**HS**	**OF**	**CMS**	**Eskin**
Snorre	39	59	39	40	34
Snohvit	57	66	15	29	27

Table 4: Mean Average Precision

	ICS	**HS**	**OF**	**CMS**	**Eskin**
MAP	48	63	27	35	30

Usually, the performance of a recommendation system is calculated by considering a set of queries. Therefore, given a set of queries Q, the mean average precision($MAP_Q@K$) of an algorithm is defined as

$$MAP_Q@K = \frac{1}{Q} \sum_{q=1}^{Q} AP_q@K \tag{11}$$

where $AP_q@K$ is calculated by using Equation 10

4.7 Experimental Results

There is no standard way to evaluate similarity measures for semantic similarity. Resnik uses human expert similarity ranking to judge similarity (Resnik 1970). We follow the same approach. In order to perform this evaluation, we selected two target reservoirs 'Snorre' and 'Snøhvit.' We then asked our domain experts to produce a gold set for each reservoir. This gold set contains a set of reservoirs identified by our experts as most similar to the target reservoir based on their hindsight knowledge about the target reservoir. Furthermore, the gold set is produced in a ranked manner, the first item in the list corresponds to the highest similar analogue and the last item corresponds to the lowest similar reservoir.

After acquiring the gold dataset, we perform an experimental evaluation to compare the performance of the proposed techniques with three existing similarity measures (OF (Boriah, Chandola, and Kumar 2008) , Eskin (Eskin et al. 2002) , CMS (Jian et al. 2018) for finding reservoir analogues. For each selected target reservoir, all the remaining reservoirs in the dataset are given as input to each similarity measure and the similarity between the target and all remaining reservoirs is calculated. The top 15 reservoirs with maximum similarity are retrieved and are now referred to as analogues to the target reservoir.

In order to penalize poor estimations, we are using Average Precision (e Equation 10) as a quality criterion for evaluation of similarity between reservoirs. For this metric, a higher value corresponds to better results. Table 3, shows the experimental result of each similarity measure separately for each target reservoir [6].

As shown in table 3, ICS and HS measures outperform the data-driven similarity measures for both selected reservoirs. For the target reservoirs, 'Snorre' and 'Snohvit', the

[6] Similarity measure proposed by Lin (Lin 1998) doesn't retrieve any similar analogues in the top k-recommendations. Therefore, results are not included in table 3.

average precision for ICS is 39% and 57% which is higher than the average precision of other similarity measures. For HS average precision for 'Snorre' and 'Snohvit' is 59% and 66%.Further, table 4 shows that the MAP (Equation 11) for ICS and HS is 48% and 63% respectively, which significantly better than the MAP values of other algorithms. This evaluation supports the initial hypothesis that by adding domain information to the similarity measure, we can increase the similarity performance for the complex categorical data.

It is important to note that results obtained using ICS and HS are not directly comparable with the gold set provided by human experts. In order to produce a gold set, human experts take into account the geological history of the current basin, analysis of geological time periods and overall processes of formation of reservoir rocks. Furthermore, they also use conceptual facies models, reservoir simulation models, core samples and well logs for selecting appropriate analogues. In contrast to this, our experimental evaluation of the proposed technique is based only on a limited part of this information. Achieving 63% precision in this scenario highlights the fact that it is highly remarkable to correctly retrieve analogues in the top 15 recommendations based only on hierarchy-based semantic measure.

5 Conclusion & Future Work

Computing similarity measure in an unsupervised setting is a complex task. In this paper, we propose a method based on domain information extracted in the form of is-a links from a concept hierarchy. The experimental results in the previous section, show that by using domain information, the results are significantly better than the traditional methods of finding similarity only based on frequency match/mismatch. In our current work, we approach the problem by considering the lowest common ancestor in the concept hierarchy by considering mono-hierarchies only and in an unsupervised setting. In the future, we want to extend the notion of similarity for categorical data in a supervised setting for complex use cases such as mortality prediction in the medical domain. Furthermore, the idea can be extended to find similarity for categorical data in poly-hierarchies (i.e. not tree-shaped).

References

Ahmad, A., and Dey, L. 2007. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters* 28:110–118.

Alamuri, M.; Surampudi, B.; and Negi, A. 2014. A survey of distance/similarity measures for categorical data. *Proceedings of the International Joint Conference on Neural Networks* 1907–1914.

Alkharusi, H. 2012. Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education* 4:202–210.

Baeza-Yates, R., and Ribeiro-Neto, B. 2008. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison-Wesley Publishing Company, 2nd edition.

Boriah, S.; Chandola, V.; and Kumar, V. 2008. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the SIAM International Conference on Data Mining*, volume 30, 243–254.

Cost, S., and Salzberg, S. 1998. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* 10.

Das, G., and Mannila, H. 2000. Context-based similarity measures for categorical databases. In *Lecture Notes in Computer Science*, volume 1910, 201–210.

Davis, M. J. 2010. Contrast coding in multiple regression analysis: Strengths, weaknesses, and utility of popular coding structures. In *Journal of Data Science*.

Eskin, E.; Arnold, A.; Prerau, M.; Portnoy, L.; and Stolfo, S. 2002. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *Applications of Data Mining in Computer Security* 6.

Esposito, F.; Malerba, D.; Tamma, V.; and Bock, H.-H. 2000. *Classical resemblance measures*. Springer Verlag.

Gambaryan, P. 1964. A mathematical model for taxonomy. *SSR* 47–53.

Huang, G., and Sheng, J. 2012. Measuring similarity between sentence fragments. In *Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2012*, volume 1, 327–330.

Jian, S.; Cao, L.; Lu, K.; and Gao, H. 2018. Unsupervised coupled metric similarity for non-iid categorical data. *IEEE Transactions on Knowledge and Data Engineering* PP:1–1.

Leacock, C., and Chodorow, M. 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, volume 49. MIT Press.

Lin, D. 1998. An information-theoretic definition of similarity. *ICML. Madison* 1.

Liu Ling, Özsu, M. T. 2009. *Encyclopedia of Database Systems*. Springer US.

Martín Rodríguez, H.; Escobar, E.; Embid, S.; Rodriguez, N.; Hegazy, M.; and Lake, L. 2013. New approach to identify analogue reservoirs. *SPE Economics & Management* 6.

Pedersen, T.; Pakhomov, S.; and Patwardhan, S. 2005. Measures of semantic similarity and relatedness in the medical domain. *Journal of Biomedical Informatics - JBI*.

Perez-Valiente, M.; Rodriguez, H.; Santos, C.; Vieira, M.; and Embid, S. 2014. Identification of reservoir analogues in the presence of uncertainty. *SPE Intelligent Energy Conference and Exhibition*.

Resnik, P. 1970. Using information content to evaluate semantic similarity in a taxonomy. *IJCAI* 95.

Ross, S. M. 1976. *A First Course in Probability*. Pearson Education, Inc.

Roy, R.; Hafedh, M.; Ellen, B.; and Maria, B. 1989. "development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19:17–30.

Smelser, N., and Baltes, P. 2001. *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier.

Sparck Jones, K. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28:493–502.

Stanfill, C., and L. Waltz, D. 1986. Toward memory-based reasoning. *Commun. ACM* 29:1213–1228.

W. Goodall, D. 1966. A new similarity index based on probability. *Biometrics* 22.

Wilson, D., and Martinez, T. 2000. Improved heterogeneous distance functions. *J. of Artif. Intell. Res.* 6.