ENTROPY ON ONTOLOGY AND INDEXING IN INFORMATION RETRIEVAL

Yevgeniy Guseynov
Optimal Solutions & Technologies, 2001 M Street, NW, Suite 3000, Washington, DC 20036, U.S.A.

Keywords: Foundation of information retrieval, Indexing in information retrieval.

Abstract: In this paper, we present a formalization of an Index Assignment process that was used against documents stored in a text database. The process uses key phrases or terms from a hierarchical thesaurus or ontology and is based on the new notion of entropy on ontology for terms and their weights that is an extension of the Shannon concept of entropy in Information Theory and the Resnik semantic similarity measure for terms on ontology. Introduced notion provides a measure of closeness or semantic similarity for a set of terms in ontology and their weights and allows creation of a clustering algorithm that constructively resolves index assignment task. The algorithm was tested on 30,000 documents randomly extracted from MEDLINE biomedicine database that are manually indexed by professional indexers. The main output from experiments shows that after all 30,000 documents were processed in seven topics out of ten the presented algorithm and human indexers have the same understanding of documents.

1 INTRODUCTION

Over past decades many Information Retrieval (IR) Systems were developed to manage the increasing complexity of textual (document) databases, see references in Manning, Raghavan, Schütze (2008). Many of these systems use a knowledge base, such as a hierarchical Indexing Thesaurus or Ontology to extract, represent, store, and retrieve information that describes such documents (Salton, 1989; Agrawal, Chakrabarti, Dom, Raghavan, 2001; Tudhope, Alani, Jones, 2001; Aronson, Mork, Gay, Humphrey, Rogers, 2004; Medelyan, Witten, 2006a; Wolfram, Zhang, 2008; and others). Ontologies were used in IR systems to endorse the semantic concepts consistency and enhance the search capabilities. In this paper we assume that ontology has hierarchical relations among concepts and interchangeably refer to ontology as hierarchical Indexing Thesaurus (Cho, Choi, Kim, Park, Kim, 2007). An Indexing Thesaurus consists of terms (words or phrases) describing concepts in documents that are arranged in a hierarchy and have a stated relations such as synonyms, associations, or hierarchical relationships among them. We discuss this in more detail later in the "Knowledge Base" section. Medical Subject Headings (MeSH) hierarchical thesaurus (Nelson, Johnston, Humphreys, 2001) together with the National Library of Medicine MEDLINE® database and the Unified Medical Language System Knowledge Source (Lindberg, Humphreys, McCray, 1993) are the best examples of IR systems for biomedical information.

There are numerous ontologies available for linguistic or IR purposes, see references in Grobelnik, Brank, Fortuna, Mozetič (2008). Mostly, they were manually built and maintained over the years by human editors (Nelson et al., 2001). There were also attempts to generate ontologies automatically by using the word's co-occurrence in a corpus of texts (Qiu, Frei, 1993; Schütze, 1998).

It is an issue in linguistics to determine what a word is and what a phrase is (Manning, Schütze, 1999). We use terminology from the Stanford Statistical Parser (Klein, Manning, 2003) which for a given text specifies part-of-speech tagged text, sentence structure trees, and grammatical relations between different parts of sentences. This information allows us to construct a list of terms from a given ontology to be used to present the initial text.

To retrieve information from databases, documents are usually indexed using terms from ontologies or key phrases extracted from the text based on their frequency or length. Indexing based

Guseynov Y..
ENTROPY ON ONTOLOGY AND INDEXING IN INFORMATION RETRIEVAL.
DOI: 10.5220/0003298205550567
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 555-567
ISBN: 978-989-8425-51-5

on ontology is typically a manual or semi-automated process that is aided by a computer system to produce recommended indexing terms (Aronson et al., 2004). For large textual databases, manual Index Assignment is highly labor-intensive process, and moreover, it cannot be consistent because it reflects the interpretations of many different indexers involved in the process (Rolling, 1981; Medelyan, Witten, 2006b). Another problem is the natural evolution of the indexing thesauruses when new terms have to be added or when some terms become obsolete. This also adds inconsistency to the indexing process. These two significant setbacks drove the development of different techniques for automating Index Assignment, see references in Manning et al. (2008), Medelyan, Witten (2006a) but none of them could be close in comparison with Index Assignment by professional indexers. Névéol, Shooshan, Humphrey, Mork, Aronson (2009) described the challenging aspects of automatic indexing using a large controlled vocabulary, and also provided a comprehensive review of work on indexing in the biomedical domain.

This paper presents a new formal approach to the Index Assignment process that uses key phrases or terms from a hierarchical thesaurus or ontology. This process is based on the new notion of Entropy on Ontology for terms and their weights and is an extension of the Shannon (1948) concept of entropy in Information Theory and the Resnik (1995) semantic similarity measure for terms in ontology. This notion of entropy provides a measure of closeness or semantic similarity for a set of terms in ontology and their weights, and is used to define the best or optimal estimation for the State of the Document, which is a pair of terms and weights that internally describes main topics in the document. This similarity measure for terms allows the creation of a clustering algorithm to build a close estimation of the State of the Document and constructively resolve Index Assignment task. This algorithm, as a main part of Automated Index Assignment System (AIAS), was tested on 30,000 documents randomly extracted from MEDLINE biomedicine database. All MEDLINE documents are manually indexed by professional indexers and terms assigned by AIAS were compared against human choices. The main output from our experiments shows that after all 30,000 documents were processed, in seven out of ten topics, AIAS and human indexers had the same understanding of the documents.

Every document in a database has some internal meaning. We may present this meaning by using a set of terms $\{T_k\}$ from the Indexing Thesaurus and their weights $\{W_k\}$ showing the relative importance of corresponding terms. We define the State of the Document as a latent pair $(\{T_k\}, \{W_k\})$ that represents implicit internal meaning of the document. The goal in Index Assignment in IR is to classify the main topics of the document to identify its state. Usually, the State of the Document is unknown, and we may have only a certain estimation of it. Among human estimations we have the following:

1. The author's estimation – how author of the document desires to see it;
2. The indexer's estimation – with general knowledge of the subject and available vocabulary from Indexing Thesaurus;
3. The user's estimation – with the knowledge of the specific field.

In addition, inside each human category the choice of the terms depends on background, education, and other skills that different readers may have and this adds inconsistency in the indexing process as mentioned earlier.

One of the thesaurus-based algorithms exploiting semantic word disambiguation was proposed in Walker (1987). The main idea here is, for a given word from the text that corresponds to different terms in thesaurus hierarchy, to choose the term *T* having the highest sum of occurrences or the highest concentration of words from the document with highest frequencies in sub hierarchy with the root *T*.

In another thesaurus based algorithm (Medelyan, Witten, 2006a) the idea of concentration based on the number of thesaurus links that connect candidate terms was mentioned as one of the useful features in assigning key phrases to the document. The same idea of word concentration that is used to identify topics or terms in the document is implicitly seen in Figure 1. Figure 1 demonstrates part of the MeSH hierarchy (Nelson et al., 2001) and MeSH terms, indicated as ▲, that were manually chosen by a MEDLINE indexer for the abstract from MEDLINE database presented in Appendix A. The MeSH terms that have a word from this abstract are spread among MeSH hierarchy in almost 30 top topics, not all of them are shown here. However, only terms that are concentrated in two related topics in ontology (hierarchy) with highest word frequencies were chosen by the indexer: "Nursing", hierarchy code G02.478, and "Health Services Administration", hierarchy code N04.

We might emphasize two main concepts that could indicate how the terms were chosen among all possible candidates in these examples: the concept of relevant or similar terms in an ontology and the

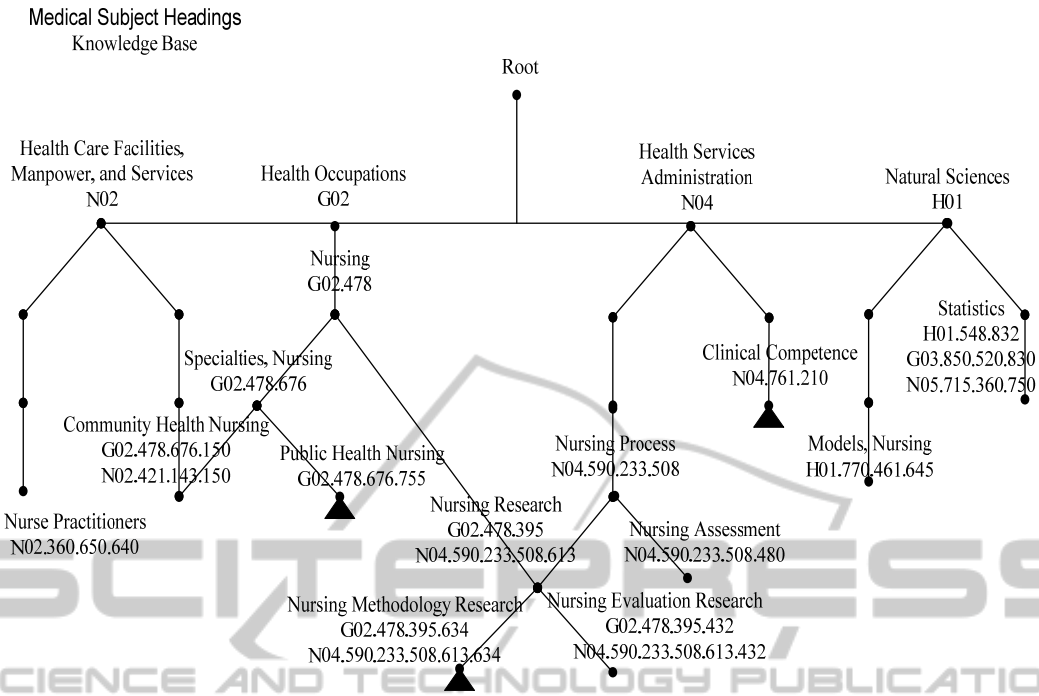

Figure 1: Medical Subject Headings for MEDLINE abstract 21116432 "Expert public health nursing practice: a complex tapestry".

concept of concentration of relevant terms in ontology that have the highest frequencies of words from the document.

The notion of concentration energy, information, business and other entities was defined through concept of Entropy (Wiener, 1961). Shannon (1948) presented the concept of Entropy in Information Theory as

$$H(\{x_k\}) = -\sum_k x_k \log_2 x_k ,$$

where $\{x_k\}$ is a distribution, $x_k \geq 0, \sum_k x_k = 1$. The functional $H(\{x_k\})$ is widely used to measure information in a distribution, particularly, to compare ontologies (Cho et al., 2007) and to measure distance between two concepts in ontology (Calmet, Daemi, 2004). Functional $H(\{x_k\})$, or entropy, is at its maximum when all $x_k, k = 1, \dots, K$, are equal, meaning that we cannot accentuate any element or a group of elements in the distribution, or, in other words, there is no concentration of information. On the other hand, functional $H(\{x_k\})$, or entropy, is 0 or at its minimum if one of the elements, say $x_1 = 1$, and all the others are 0. In this case all information about the distribution is well known and is concentrated in x_1.

The concept of similarity for two terms in IS-A ontology was introduced by Resnik (1995, 1999) and is based on the information content -$\log_2 p(T)$, where $p(T)$) is an empirical probability function of terms T in ontology. The measure of similarity for terms T_1 and T_2 is defined as the maximum information content evaluated over all terms that subsume both T_1 and T_2. The measure of similarity is used in linguistics, biology, psychology and other fields to find semantic relationships among the entities of ontologies (Resnik, 1999).

In IR, the input set of weights $\{W_k\}$ for candidate terms is usually not a distribution, and so we extend the concept of entropy for weights $W_k > 0$, $\sum_k W_k \neq 1$. We also expand the concept of similarity to measure the similarity for any set of terms in ontology. Based on these new notions of Weight Entropy and Semantic Similarity, we introduce in corresponding section the notion of Entropy on Ontology for any set of candidate terms $\{T_k\}$ and their weights $\{W_k\}$. We define Optimal Estimation of the State of the Document as a pair $(\{T_k\}, \{W_k\})$ where the minimum value for Entropy on Ontology is attained over all possible sets of candidate terms. Theoretically, this is a formal solution for the Index Assignment problem and the minimum of entropy could be found through enumeration of all possible cases. Compared to human indexers, the Optimal Estimation of the State of the Document provides a uniform approach to solving the problem of assigning indexing terms to

documents with the vocabulary from an indexing thesaurus. Any hierarchical knowledge base can be used as an indexing thesaurus for any businesses, educational, or governmental institutions.

In general, when the indexing thesaurus is too large, the Optimal Estimation of the State of the Document provides a non-constructive solution to the problem of assigning indexing terms to a document, see also Névéol et al. (2009) for the scalability issue. Nevertheless, its definition provides insight into how to construct a Quasi Optimal Estimation that is presented in corresponding section. We may consider the Index Assignment problem as a process that is used to comprehend and cluster all possible candidate terms with words from the given document into groups of related terms from the indexing thesaurus that present the main topics of the document. There are different clustering algorithms, particularly in IR (Manning, Schütze, 1999; Rasmussen, 1992), that characterize the objects into groups according to predefined rules that represent formalized concepts of similarity or closeness between objects. Rather than randomly enumerating all possible sets of candidate terms, we use clustering. We start from separate clusters for each term that contains a word from a given document to construct a Quasi Optimal Estimation algorithm for the State of the Document. It is based on the concept of closeness introduced here as Entropy on Ontology which evaluates similarity for a set of terms and their weights.

Manually indexed documents are the best candidates for testing the new algorithm, and maybe unique samples for comparison in assignment terms from ontology. We evaluate our algorithm against human indexing of abstracts (documents) from MEDLINE bibliographic database covering the fields with concentration on biomedicine. MEDLINE contains over 16 million references to journal articles in life sciences worldwide and over 500,000 references are added every year. A distinctive feature of MEDLINE is that the records are indexed with Medical Subject Headings (MeSH) Knowledge Base (Nelson et al., 2001) which has over 25,000 terms and 11 levels of hierarchy. The evaluation results are discussed in "Algorithm Evaluation" section.

2 KNOWLEDGE BASE

The knowledge base for any domain of the world and any human activity consists of semantic interpretation of words from documents that may be used for indexing. One of the organizations of such knowledge is Hierarchical Indexing Thesaurus or Ontology. Terms for ontology are usually selected and extracted based on the users' terminology or key phrases found in documents stored in the database. Each term should represent a topic or a feature of the knowledge domain and provide the means for searching the database for this topic or feature in a unique manner.

The other fundamental components in ontology are hierarchical, equivalence, and associative relationships.

The main hierarchical relationships are:

part/whole, where relation may be described as "A is part of B", "B consists of";

class/subclass, where child term inherits all features of the parent and has its own properties;

class/object, where the term A as an object is instantiated based on the given class B, and "A is defined by B".

Equivalence in relationships may be described also as "term A is term B", when the same term is applied to two or more hierarchical branches, as in the most concerned situation.

Associative relationship is a type of "see related" or "see also" cross-reference. It shows that there is another term in the thesaurus that is relevant and should also be considered.

Two terms in ontology may relate to each other in other ways. A concept of similarity that measures relationship between two terms was introduced by Resnik (1995) and is based on the prior probability function $p(T)$ of encountering term T in documents from a corpus. This probability function can be estimated using the frequencies of terms from the corpora (Resnik, 1995; Manning, Schütze, 1999). A formal definition that will be used in the sequel is as follows:

An Ontology or a Hierarchical Indexing Thesaurus is an acyclic graph with hierarchical relationships described above, together with a prior probability function $p(T)$ that is monotonic: if term T_1 is a parent of term T, then $p(T) \leq p(T_1)$ (Resnik, 1995); in case of multiple parents and if the number of parents equals N_T we will assume that $\frac{\mathrm{p}(T)}{N_T} \leq \mathrm{p}(T_1)$ for each parent T_1. Nodes on the graph are labeled with words or phrases from the documents' database. The graph has a root node called "Root", with p(Root) = 1. All other nodes have at least one parent. Some nodes may have multiple parents, which represent the equivalence or associative relationships between nodes. Figure 1 shows an example of an acyclic graph from the MeSH Indexing Thesaurus.

3 ENTROPY ON ONTOLOGY

The State of the Document, as defined in the introduction, is a set of terms from ontology with weights that provides an implicit semantic meaning of the document, and, in most cases, is unknown. Having multiple estimations of the State of the Document, we need to have a measurement that would allow us to distinguish different estimations in order to find the one most closely describing the document.

3.1 Weight Entropy

Examples discussed in the introduction (Walker, 1987; Medelyan, Witten, 2006a; Nelson et al., 2001) demonstrate the importance of measuring the concentration of information presented in a set of weights and the entropy $H(\{x_k\})$ (Shannon, 1948) for a distribution $\{x_k\}, x_k \geq 0, \sum_k x_k = 1$, is a unique such measurement. In IR, the input set of weights $\{W_k\}$ is usually not a distribution and replacement of weights with normalized weights $\left\{\frac{W_k}{\sum_k W_k}\right\}$, $\sum_k W_k \neq 0$, leads to a loss of several important weight features. Intuitively, when sum $\sum_k W_k \to 0$, the weights vanish and provide less substance for consideration, or less information. Similarly, if we have two sets of weights with the same distribution after normalization, we cannot distinguish them based on normalized weights and classical entropy *H*. However, one of the weights' sums could be much bigger than the other and we should choose first one as an estimation of the State of the Document. Also, in the simplest situation, when we want to compare sets each of which consists of just one term, all normalized weights will have zero entropy *H,* and again, the term with bigger weight would be preferable. After these simple considerations, we define Weight Entropy for weights $\sum_k W_k \neq 0$ as

$$WE(\{W_k\}) = \frac{1}{\sum_k W_k}\left(1 + \mathrm{H}\left(\left\{\frac{W_k}{\sum_k W_k}\right\}\right)\right)$$

$$= \frac{1}{\sum_k W_k}\left(1 - \sum_k \frac{W_k}{\sum_k W_k}\log_2\frac{W_k}{\sum_k W_k}\right).$$

As we see from the definition, in addition to the features of classic entropy, this formula allows us to utilize the substance of the sum of the weights when comparing sets of weights. We also see that for $\sum_k W_k = 1$ we have

$$WE(\{W_k\}) = 1 + H(\{W_k\})|_{\sum_k W_k = 1},$$

and so in this case the weight entropy is classic entropy plus 1. Slight modification of the definition of $WE(\{W_k\})$ would result in

$$W(\{W_k\}) = H(\{W_k\})|_{\sum_k W_k = 1},$$

but this is not important for our considerations below.

3.2 Semantic Similarity

Semantic similarity is another important concept emphasized in the introduction. Let's assume that we evaluate semantic similarity between two sets of terms in ontology presented in Figure 1. Let set S_1 = {"Community Health Nursing", "Nursing Research", "Nursing Assessment"} = $\{T_1, T_2, T_3\}$. We would like to compare S_1 with the set S_2 = $\{T_1, T_2, T_4\}$, where T_4= "Clinical Competence"; we want to focus only on the topologies of sets S_1 and S_2 in ontology without weights. Empirically, we may be able to tell that the terms in set S_1 are much more similar in a given ontology than the terms in set S_2. We may also evaluate the level of similarity based on the Similarity Measure (Resnik, 1995) or the Edge Counting Metric (Lee, Kim, Lee, 1993) to formally prove our empirical choice.

In general, we compose a Semantic Similarity Cover for set *S* by constructing a set *ST* of sub trees in ontology which have all their elements from *S*. Only these elements are leaf nodes and each two nodes from *S* have a path of links leading from one node to another; all are in *ST*. We can always do this because each node from the ontology has a (grand) parent as a root node. If a sub tree from *ST* has a root node as an element, we can try to construct another extension for set *S* to exclude the root node. If at least one such continuation does not have a root node as an element, we say that set *S* has a semantic similarity cover *SSC*(*S*), or that the elements of set *S* are semantically related. If *S* cannot be extended to *SSC*, let *SP* = $\{S_i\}$ be a partition of *S,* where each set S_i has *SSC*(S_i). Some of S_i may consist of only one term. In this case S_i itself would be *SSC* for S_i . We may assume that $SSC(S_i) \cap SSC(S_j) = \emptyset$ for $i \neq j$. We say that set *S* consists of semantically related terms, or is semantically related, if set *S* has a semantic similarity cover *SSC*(*S*). Below we list several properties of semantic similarity that we will use, such as:

If set *S* is semantically related, then any cover *SSC*(*S*) is semantically related.

If set S_1 is semantically related to S_2, i.e. $S_1 \cup S_2$ is semantically related, then *SSC*(S_1) is semantically related to *SSC*(S_2).

If set S_1 is semantically related to S_2, then for each cover $SSC(S_1)$ and $SSC(S_2)$, $SSC(S_1) \cap SSC(S_2) = \emptyset$, there are $T_1 \in \mathrm{SSC}(S_1)$ and $T_2 \in \mathrm{SSC}(S_2)$ that are semantically related. Thus, they have a common parent that is not Root. (See proof in Appendix B).

To measure semantic similarity for terms from set S in ontology with prior probability function p we define the following:

$$sim(S) = max_{SSC(S)} min_{T \in SSC(S)}(-\log_2 p(T)),$$

where max is taken over all semantic similarity covers SSC(S). If S does not have SSC then we put sim(S) = 0.

The notion SSC for a set S is a generalization of Resnik's construction of semantic similarity for a pair of terms and $sim(S)$ for the pairs equals the similarity measure introduced in (Resnik, 1995). It is not used in further constructions and is presented here only for comparison purposes.

3.3 Weight Extension

Finally, we need to extend the initial weights $\{W_k\}$ for semantically related terms $\{T_k\}$ over $SSC(\{T_k\})$ using the prior probability function p. This will allow us to view SSC as a connected component and involve ontology as a topological space in the entropy definition. We assign a posterior weight continuation value $PW(T)$ for each term T from $SSC(\{T_k\})$ starting from leaf terms that are all from $\{T_k\}$ by construction. We would like to carry on value of $\sum_k W_k$ for extension to maintain important features of weights.

For each leaf term T where $T = T_k$, we define the initial weight $IW(T) = W_k$.

Using $IW(T)$ we recursively define a posterior weight $PW(T)$ for each term T, starting from the leaf terms and a posterior weight $PW(P \mid T)$ for all parents of term T from $SSC(\{T_k\})$.

If term T does not have a parent from $SSC(\{T_k\})$, we define posterior weight $PW(T) = IW(T)$ and move to the next term from current level.

Let N_T be the number of parents and let $P_1, \ldots, P_N, N \le N_T$, be the parents from $\mathrm{SSC}(\{T_k\})$ for term T with the defined initial weight $IW(T)$. We define posterior weights as

$$PW(T) = IW(T)\left(1 - \frac{1}{N\,N_T}\sum_{n=1}^{N}\frac{p(T)}{p(P_n)}\right),$$

$$PW(P_n\,|T) = IW(T)\frac{p(T)}{N\,N_T\,p(P_n)}, n = 1, \ldots N$$

Particularly, these formulas define $PW(T)$ for all leaf terms T and $PW(P|T)$ for all their parents from $SSC(\{T_k\})$ for the first level. Now we may move from leaf terms to the next level which is a set of their parents in $SSC(\{T_k\})$, to define posterior weights.

Let T be equal to a term from $SSC(\{T_k\})$ for which we have $PW(T|C)$ for all its children C from $SSC(\{T_k\})$. For this term we define initial weight as:

$$IW(T) = W_k + \sum_{C \in SSC(\{T_k\}) \cap Children(T)} PW(T|C),$$

if T is one of $T_k \in \{T_k\}$, otherwise

$$IW(T) = \sum_{C \in SSC(\{T_k\}) \cap Children(T)} PW(T|C),$$

where $Children(T)$ is the set of all children of node T.

If term T does not have a parent in $SSC(\{T_k\})$, then we define $PW(T) = IW(T)$ and the process stops for the branch with the root T. Otherwise we have to recursively calculate the posterior weights $PW(T)$ and $PW(P \mid T)$ like we did earlier for T and all its parents P derived from $SSC(\{T_k\})$.

The process should continue until the weight continuation value $PW(T)$ is defined for all terms from $SSC(\{T_k\})$ and by construction

$$\sum_{T \in SSC(\{T_k\})} PW(T) = \sum_k W_k.$$

Now we can define Entropy on Ontology with the prior probability function p for any pair of ($\{T_k\}$, $\{W_k\}$), when $\{T_k\}$ are semantically related terms, as:

$$EO(\{T_k\},\{W_k\}) = \min_{\mathrm{SSC}(\{T_k\})} WE\left(\{PW(T)\}_{T \in \mathrm{SSC}(\{T_k\})}\right) =$$

$$= \min_{\mathrm{SSC}(\{T_k\})} \frac{1}{\sum_k W_k} \times$$

$$\left(1 - \sum_{T \in \mathrm{SSC}(\{T_k\})} \frac{PW(T)}{\sum_k W_k}\log_2 \frac{PW(T)}{\sum_k W_k}\right),$$

where $PW(T)$ is the weight extension for T, and the minimum is taken over all possible covers $SSC(\{T_k\})$.

4 OPTIMAL ESTIMATION OF THE STATE OF THE DOCUMENT

The above defined notion of Entropy on Ontology (EO) provides an efficient way to measure the semantic similarity between given terms with posterior weights. It allows us to evaluate different estimations and define the best one or the optimal one for this measurement.

When we process a document *D*, we observe words with their frequencies $\{Q_i\}$. This observation gives us a set of terms *S*(*D*) from ontology that have one or more words from this document. The observation weight *W* of the term *T* that has a word from the document is calculated based on the words' frequencies $\{Q_i\}$. For example, if we want to take into consideration not only frequencies but also how many words from a document are presented in term *T* we would use the following:

$$W(T) = \left(m\frac{m}{h} + 1\right)\sum_{i=1}^{m} Q_i,$$

where *h* is number of words in *T* and words $w_1, \dots, w_m$, m ≤ h, from *T* have positive frequencies $Q_1, \dots, Q_m$. The observation weight *W* could be more sophisticated. In addition, we may set *W*(*T*) = 0 if some specific word from term *T* is not presented in the document. We leave this discussion for the next publication where we present document processing using our Automated Index Assignment System. For now we need to know that for each term *T* and given frequencies $\{Q_i\}$ of words from *D* we can calculate observation or posterior weight *W* = *W*(*T*).

Some words may participate in many terms from set *S*(*D*)= $\{T_k\}$. Let $\{R_{ik}\}$ be a partition of $\{Q_i\}$ among terms $\{T_k\}$ with $\sum_k R_{ik} = Q_i, Q_i \in \{Q_i\}$. There may be many different partitions $\{R_{ik}\}$ of words frequencies $\{Q_i\}$ for document *D*. For each partition we calculate the observation, or posterior, set of weights $\{W(T,\{R_{ik}\})\}_{T\epsilon\, S(D)}$ to find out how words from the document should be distributed among the terms in order to define its state.

For each partition $\{R_{im}\}$ we consider a set of terms $\{T_k\}$ and their weights $\{W_k\}$, where

$W_k = W(T_k,\{R_{im}\}) > 0$. Let $\{S_j\}$ be a partition $\{T_k\}$ where each S_j has a semantic similarity cover. An optimal estimation of the State of the Document is a semantic similarity cover $\{SSC(S_j)\}$ with its posterior weights minimizing the Entropy on Ontology

$$\sum_j EO(S_j, \{W(T,\{R_{im}\})\}_{T\,\in\, S_j})$$

over all partitions $\{R_{im}\}$ of words frequencies $\{Q_i\}$ among terms *S*(*D*) with $\sum_m R_{im} = Q_i, Q_i \in \{Q_i\}$ and over all partitions $\{S_j\}$ for set of terms *T* where $W(T,\{R_{im}\}) > 0$ and S_j consists of semantically related terms. Last partition $\{S_j\}$ we need not only to split by sets that consist of semantically related terms. This also allows to discover different semantic topics that may be presented in a document even if they are semantically related.

We can rewrite the optimal estimation of the state in terms of the functional:

$$G(\{R_{im}\}, S_j, \mathrm{SSC}(S_j)) = WE\left(\{PW(T)\}_{T\,\in\,\mathrm{SSC}(S_j)}\right)$$

$$= \frac{1}{\sum_{T\in S_j} W(T,\{R_{im}\})}\,(1 - \sum_{T\in \mathrm{SSC}(S_j)} \frac{PW(T)}{\sum_{T\in S_j} W(T,\{R_{im}\})} \log_2 \frac{PW(T)}{\sum_{T\in S_j} W(T,\{R_{im}\})})$$

by adding parameter $\{SSC(S_j)\}$ to the minimization area that was hidden in the definition of the Entropy on Ontology.

Finding the minimum of such a functional to construct the optimal estimation for documents from a database and a large ontology is still challenging for mathematicians and instead, below we consider a quasi solution.

5 QUASI OPTIMAL ESTIMATION ALGORITHM

Functional *G,* introduced in the previous section, provides a metric defined by $\{R_{jk}\}$, S_i, and $SSC(S_i)$ to evaluate what group of terms and their weights are closer to optimal estimation and therefore have less Entropy on Ontology or are more informative compared with others. Based on this metric we will use a "greedy" clustering algorithm that defines groups or clusters $\{S_i\}$ for set of terms *S*(*D*), where the observation weight *W*(*T*) > 0, words distribution $\{R_{jk}\}$ among the terms inside each group S_i, a cover $\mathrm{SSC}(S_i)$, that for each step creates an approximation to the optimal estimation of the State of the Document.

1. We start from separate cluster for each term from *S*(*D*) and calculate functional *G* for each $S_i = \{T\}$, $T \in S(D)$. Set S_i consists of one term *T* and $\{R_{jk}\}$ would be $\{Q_j\}$, because there is no

need to have a partition for cluster T and $SSC(S_i) = \{T\}$. Having these in place we see that for each $T \in S(D)$ $G(\{R_{jk}\}, S_i, SSC(S_i)) = 1 / W(T)$.

2. Recursively, we assume that we have completed several levels and obtained set $C_1 = \{S_i\}$ of clusters with defined word distribution $\{R_{jk}(S_i)\}$ and a semantic similarity cover $SSC(S_i)$ for each $S_i \in C_1$ where $\{R_{jk}(S_i)\}$ provides the minimum for $G(\{R_{jk}\}, S_i, SSC(S_i))$ over all partitions $\{R_{jk}\}$ of words frequencies $\{Q_j\}$.
3. In the next level we try to cluster each pair $S_1, S_2 \in C_1, S_1 \neq S_2$ that are semantically related to low values $G(\{R_{jk}\}, S_i, \mathrm{SSC}(S_i))$, $i = 1, 2$.
   3.1. New cluster construction.
      3.1.1. If semantic similarity covers for S_1 and S_2 have common terms we choose $SSC(S_1) \cup SSC(S_2)$ as the semantic similarity cover for $S_1 \cup S_2$ and recalculate $\{R_{jk}(S_1 \cup S_2)\}$ to minimize $G(\{R_{jk}\}, S_1 \cup S_2, SSC(S_1) \cup SSC(S_2))$ over all partitions $\{R_{jk}\}$.
      3.1.2. If the semantic similarity covers $SSC(S_1)$ and $SSC(S_2)$ do not have common terms, we first construct $SSC(S_1 \cup S_2)$ using the semantically related pairs of $T_1 \in \mathrm{SSC}(S_1)$ and $T_2 \in \mathrm{SSC}(S_2)$ with common parent that we know exist. For pair $\{T_1, T_2\}$ consider set $P(\{T_1, T_2\})$ of all of the closest parents, i.e. parents that do not have other common parent for T_1 and T_2 on their branch. For each such parent we may construct $SSC(\{T_1, T_2\})$ and evaluate $G(\{R_{jk}\}, S_1 \cup S_2, SSC(S_1) \cup SSC(S_2) \cup SSC(\{T_1, T_2\})$. The cover $SSC(S_1) \cup SSC(S_2) \cup SSC(\{T_1, T_2\})$ and partition $\{R_{jk}(S_1 \cup S_2)\}$ that provide minimum for G over all such covers $SSC(\{T_1, T_2\})$ and partitions $\{R_{jk}\}$ is the new cluster for $S_1 \cup S_2$.
   3.2. Let cluster $S_1 \in C_1$ have the minimum value for functional G over all clusters in C_1 and $C_2 = C_1$.
      3.2.1. For each cluster $S_2 \in C_1$, $S_1 \neq S_2$, semantically related $to\ S_1$, we construct new $SSC(S_1 \cup S_2)$ like it is done in 3.1. If $G(\{R_{jk}(S_1)\}, S_1, SSC(S_1)) + G(\{R_{jk}(S_2)\}, S_2, SSC(S_2)) \geq G(\{R_{jk}(S_1 \cup S_2)\}, S_1 \cup S_2, \mathrm{SSC}(S_1) \cup SSC(S_2) \cup \mathrm{SSC}(\{T_1, T_2\})$, then we mark value $G(\{R_{jk}(S_1 \cup S_2)\}, S_1 \cup S_2, \mathrm{SSC}(S_1) \cup SSC(S_2) \cup \mathrm{SSC}(\{T_1, T_2\})$ for comparison. We exclude cluster S_2 from set C_1.
      3.2.2. We repeat step 3.2.1, until all elements from C_1 are processed, to choose a cluster with the lowest value $\mathrm{G}(\{R_{jk}(S_1 \cup S_2)\}, S_1 \cup S_2, \mathrm{SSC}(S_1) \cup SSC(S_2) \cup \mathrm{SSC}(\{T_1, T_2\})$ for joined cluster.
      3.2.3. If the cluster in 3.2.2 exists, we exclude from C_2 clusters S_1 and S_2 that we chose in step 3.2.2 and include new joined cluster $S_1 \cup S_2$ with constructed distribution $\{R_{jk}(S_1 \cup S_2)\}$, and cover $SSC(S_1 \cup S_2))$ in set C_3.
      3.2.4. If the cluster in 3.2.2 does not exist, inequality in 3.2.1 holds for any cluster $S_2 \in C_1$, we exclude cluster S_1 from C_2 and include it in C_3.
      3.2.5. We rename $C_1 = C_2$. At this point we have set C_1 reduced at least by one element and we have to go back to step 3.2 from the beginning until set C_1 is empty and set C_3 consists of clusters for the next level.
4. We rename $C_1 = C_3$. If number of clusters in newly built level C_1 did not change compared with previous level and we cannot further combine terms to reduce value of functional G, we stop the recursion process. Otherwise, we go back to step 3 to build clusters for the next level.
5. At this point we have set C_1 that consists of clusters S with defined words distribution $\{R_{jk}(S)\}$ and semantic similarity cover $SSC(S)$. Each cluster has its own words distribution and we have to construct one

distribution for the whole set C_1. We assume that in each document there is no more than one topic with the same vocabulary.

5.1. Let cluster S_1 have the lowest value of $G(\{R_{jk}\}, S_i, SSC(S_i))$ among all clusters from C_1 which indicates that cluster S_1 is the main topic in document. We exclude S_1 from C_1 and include it into final set C_2.

5.2. We exclude all frequencies of words that are part of terms from clusters in C_2 from all clusters in C_1 to create reduced set of frequencies $\{Q_j\}$. We then recalculate functional G for all clusters in C_1 based on a new set of frequencies $\{Q_j\}$, and exclude from C_1 clusters with zero values for functional G after recalculation.

5.3. Now set C_1 is reduced at least by one element and we have to repeat steps 5.1 and 5.2 until set C_1 is empty and C_2 contains the final set of clusters.

The final set of clusters, their semantic similarity covers, and the distribution of words built in steps 1 - 5, compose an approximation or Quasi Optimal Estimation for the State of the Document. The algorithm in steps 1 – 5 is one of the possible approximations to the optimal estimation of the State of the Document. It shows how semantic similarity cover and words distribution could be built recursively based on the initial frequencies of words in a document.

6 ALGORITHM EVALUATION

The implementation of the algorithm described in previous sections does not depend on any particular indexing thesaurus or ontology and can be tuned for indexing documents from any database. The algorithm is the main part of our Automated Index Assignment System (AIAS) that is very fast in processing documents based on new XML technology (Guseynov, 2009).

For algorithm evaluation we use the MeSH Indexing Thesaurus and the medical abstract database. The entire MeSH, over 25,000 records in 2008 release, was downloaded from the http://www.nlm.nih.gov/mesh/filelist.html into AIAS to build a hierarchical thesaurus for our experiments. Also the file medsamp2008a with 30,000 random MEDLINE citations was downloaded from the MEDLINE site at http://www.nlm.nih.gov/bsd/sample_records_avail.html as a sample data to be processed by our algorithm. For each MEDLINE citation (document) from this file, the estimation of State of the Document, that is the MeSH terms (MTs) and their weights, was performed. The field “Mesh Heading List” from MEDLINE citations containing terms assigned to abstracts by human MeSH Subject Analysts was used for comparison against terms assigned by our algorithm. Thus, the entire experiment was based on 30,000 documents randomly extracted from a large corpus of over 16 million abstracts, a manually built MeSH hierarchical indexing thesaurus as ontology, existing human estimation of the documents, and an estimation of the same documents produced by our algorithm.

We evaluate the algorithm based on statistics for three indicators which are similar to characteristics for the identification consistency of Index Assignments between two professional indexers (Rolling, 1981; Medelyan, Witten, 2006b).

One of the main indicators for evaluation is the ratio Matched Hierarchically. Two MeSH terms are said to be matched hierarchically, if they are on the same hierarchical branch in ontology. For example, MT “Nursing Methodology Research” with MN=G02.478.395.634 as node hierarchy and “Nursing”, MN= G02.478, are on the same hierarchical branch and are topologically close (Figure 1). The MT “Nursing Methodology Research” will always be chosen if MT “Nursing” is present. The Matched Hierarchically indicator is the ratio of the number of MTs from MEDLINE, each of which matched hierarchically to some MT chosen by AIAS, to the total number of MEDLINE terms.

The second indicator is Compare Equal. For term T_1 assigned by AIAS and term T_2 assigned by the MEDLINE indexer that are matched hierarchically we calculate the minimum number of links between them on the MeSH hierarchy. We assign a plus sign to the number of links if term T_1 is a child of T_2, and a minus sign, if T_1 is a parent of T_2. For each document the average number of signed links for matched hierarchically terms represents the Compare Equal indicator.

The third indicator is Ratio AIAS to MEDLINE Terms which for each document is the ratio of the total number of terms chosen by AIAS to the total number of terms chosen by the MEDLINE indexer.

The meanings of all these indicators are evident in vector space model for IR systems (Manning et al., 2008; Wolfram, Zhang, 2008). The most important characteristics for retrieval process are the relevance indexes to a document, the preciseness of the indexing terms, in our case the deeper in

hierarchy a term appears, the more precise the term is said to be, and the depth of indexing, representing the number of terms used to index the document. They directly affect the index storage capacity, performance, and relevance of retrieval results. We would like the Matched Hierarchically indicator to be close to 1. It is always less than or equal to 1, and the closer it is to 1 the more terms from MEDLINE will have the same topics chosen by AIAS or the AIAS and MEDLINE indexer will have a close understanding of the document. Having this indicator close to 1 also means that the terms assigned by AIAS are relevant to the documents as MEDLINE terms are proven to be most relevant to MEDLINE citations based on people judgment and extensive use in biomedical IR. We would like the Compare Equal indicator to be close to 0. Having it less than 0 means that AIAS chose more general topics to describe the document than the MEDLINE indexer did; if it is greater than 0, the AIAS choice is more elaborate which is preferable. In the latter case, terms reside deeper in MeSH hierarchy and appear rarely in MEDLINE collection with a greater influence in the choice of relevant documents in IR. It is desirable for the ratio AIAS to MEDLINE Terms to equal to 1. In this case, the AIAS and MEDLINE indexer will choose the same number of topics to describe the document that leads to the same storage capacity and performance in IR.

The averaged output statistics after the whole medsamp2008a file was processed was:

Matched Hierarchically 0.71;
Compare Equal -0.41;
Ratio AIAS to MEDLINE 2.08.

The result 0.71 for the Matched Hierarchically is very encouraging. This indicates that in general, for all 30,000 processed citations and, for more than seven MeSH terms out of ten assigned by Subject Analysts, AIAS chose the corresponding MeSH terms on the same hierarchical branch. This means that in seven cases out of ten, the AIAS and MeSH Subject Analysts had the same understanding of the documents' main topics which shows high level of relevance between them. This result also indicates that estimations of the State of the Document in general are slightly different (three out of ten) between AIAS and the Subject Analysts.

The result -0.41 for the Compare Equal indicator means that AIAS chooses more general terms on the hierarchy in comparison to the terms from MEDLINE. This means that a greater number of documents needs to be retrieved based on AIAS, and this would make it more difficult for the users to choose a relevant document. A partial explanation for this trend is that the current release of AIAS is set to pick up more general term if two candidates have the same properties. We must prove this or find a more effective explanation.

The ratio 2.08 for the AIAS to MEDLINE indicator is too high and IR system based on AIAS would need double the storage capacity and have lower performance. This ratio means that for each 10 terms chosen by MeSH Subject Analysts to describe MEDLINE citation, AIAS needs more than 20 terms to describe the same document and many of those terms could be redundant. We can easily reduce this indicator but this will affect the Matched Hierarchically statistics. This ratio is very sensitive to the internal notion of "Stop Words" that AIAS uses now and we intend to significantly change it in the next AIAS release along with the whole approach to calculation of terms weights through words frequencies for documents. This will significantly improve our output statistics.

Going back to the MEDLINE abstract used in the Introduction in Figure 1, we now can show its statistics as:

Matched Hierarchically 0.75;
Compare Equal -0.33;
Ratio AIAS to MEDLINE 1.75.

Two terms "Public Health Nursing" and "New Zealand" were chosen by Subject Analyst. These terms were also chosen by AIAS. "Nursing Research" chosen by AIAS is one level more general than "Nursing Methodology Research" which was chosen by the Subject Analyst. The term "Clinical Competence", which did not have words from the abstract, was not chosen by AIAS. In addition to these, AIAS chose "Models, Nursing", "Nursing Assessment", "Statistics", and "Nurse Practitioners". All these were summarized in the statistics above.

The terms chosen for this abstract by AIAS and the MeSH Subject Analyst show different points of view on how a document state could be estimated. It also emphasizes, once again, the necessity of a uniform approach to the Index Assignment process.

Our evaluation of the algorithm presented in this paper was founded on a human judgment which is usually considered a "gold standard". In general, it is very difficult to obtain a large set of reliable judgments for comparison and Index Assignment processes are usually evaluated with respect to their performance with particular IR system based on statistical significance of experimental results (Salton, 1991; Aronson et al., 2004). In the last few years this type of evaluations became increasingly

demanding for IR systems and we intend to conduct these experiments in the near future.

7 CONCLUSIONS

The notion of Entropy on Ontology, introduced above, involves a topology of entities in a topological space. This feature was realized through a weight extension on the semantic similarity cover as a connected component on ontology and can be used as a pattern to similarly define entropy for entities from other topological spaces to formalize some semantics like similarity, closeness, or correlation between entities. This new notion can be used to measure information in a message or collection of entities when we know weights of entities that compose a message and, in addition, how entities "semantically" relate to each other in a topological space.

The quality of the presented algorithm that allows us to estimate Entropy on Ontology and the State of the Document depends entirely on the correctness and sufficiency of the hierarchical thesaurus on which it is based. As mentioned earlier, there are many thesauruses and their maintenance and evolution are vital for the proper functioning of such algorithms. The world also has acquired a great deal of knowledge in different forms, like dictionaries, and it is very important to convert them into a hierarchy to be used for the proper interpretation of texts that contain special topics.

The minimum that defines Entropy on Ontology and the State of the Document may not be unique or there may be multiple local minima. For developing approximations it is important to find conditions on ontology or terms topology under which the minimum is unique.

Current release of AIAS uses MeSH Descriptors vocabulary and WordWeb Pro general purpose thesaurus in electronic form to select terms from ontology using words from a document. Many misunderstandings of documents by AIAS that were automatically caught were the result of insufficiencies of these sources when processing MEDLINE abstracts. The next release will integrate the whole MeSH thesaurus, Descriptors, Qualifiers, and Supplementary Concept Records, to make AIAS more educated regarding the subject of chemistry. Also, any additional thesaurus made available electronically would be integrated into AIAS.

The algorithm that was presented in Section 5 was only tested on the MEDLINE database and MeSH ontology. Its implementation does not depend on a particular indexing thesaurus or ontology and it would be interesting to try it on other existing text corpora and appropriate ontology such as WordNet (http://wordnet.princeton.edu) or others.

REFERENCES

Agrawal, R., Chakrabarti, S., Dom, B.E., Raghavan, P., 2001. Multilevel taxonomy based on features derived from training documents classification using fisher values as discrimination values. *United State Patent* 6,233,575.

Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J., 2004. The NLM indexing initiative's Medical Text Indexer, Stud Health Technol Inform 107 (Pt 1), pp. 268–272.

Calmet, J., Daemi, A., 2004. From entropy to ontology. Fourth International Symposium "From Agent Theory to Agent Implementation", R. Trappl, Ed., vol. 2, pp. 547 – 551.

Cho, M., Choi, C., Kim, W., Park, J., Kim, P., 2007. Comparing Ontologies using Entropy. 2007 International Conference on Convergence Information Technology, Korea, 873-876.

Grobelnik, M., Brank, J., Fortuna, B., Mozetič, I., 2008. Contextualizing Ontologies with OntoLight: A Pragmatic Approach. Informatica 32, 79–84.

Guseynov, Y., 2009. XML Processing. No Parsing. Proceedings WEBIST 2009 - 5th International Conference on Web Information Systems and Technologies, INSTICC, Lisbon, Portugal, pp. 81 – 84.

Klein, D., Manning, C.D., 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Lee, J.H., Kim, M.H., Lee, Y.J., 1993. Information retrieval based on conceptual distance in IS-A hierarchies. Journal of Documentation, 49(2):188-207, June.

Lindberg, D.A.B., Humphreys, B.L., McCray, A.T., 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4): 281-91.

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Manning, C. D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

Medelyan, O., Witten, I.H., 2006a. Thesaurus Based Automatic Keyphrase Indexing. *JCDL'06*, June 11–15, Chapel Hill, North Carolina, USA.

Medelyan, O., Witten, I.H., 2006b. Measuring Inter-Indexer Consistency Using a Thesaurus. *JCDL'06*, June 11–15, Chapel Hill, North Carolina, USA.

MEDLINE®, Medical Literature, Analysis, and Retrieval System Online. http://www.nlm.nih.gov/databases/databases_medline.html.

Nelson, S.J., Johnston, J., Humphreys, B.L., 2001. Relationships in Medical Subject Headings. In: Bean, Carol A.; Green, Rebecca, editors. Relationships in the organization of knowledge. New York: Kluwer Academic Publishers. p.171-184.

Névéol, A., Shooshan, S.E., Humphrey, S.M., Mork, J.G., Aronson, A.R., 2009. A recent advance in the automatic indexing of the biomedical literature. J Biomed Inform. Oct;42(5):814-23.

Qiu, Y., Frei, H.P., 1993. Concept based query expansion. In *Proc. SIGIR*, pp. 160–169. ACM Press.

Rasmussen, E., 1992. Clustering algorithms. In William B. Frakes and Ricardo Baeza-Yates (eds.), Information Retrieval, pp. 419-442. *Englewood Cliffs, NJ: Prentice Hall*.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11, 95-130.

Rolling, L., 1981. Indexing consistency, quality and efficiency. *Information Processing and Management*, 17, 69–76.

Salton, G., 1989. Automatic Text Processing. Addison-Wesley.

Salton, G., 1991. The Smart project in automatic document retrieval. In *Proc. SIGIR*, pp. 356–358. ACM Press. 173, 530

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell System Technical Journal. 27:3 pp 379-423.

Schütze, H., 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–124.

Tudhope, D., Alani, H., Jones, C., 2001. Augmenting thesaurus relationships: possibilities for retrieval. *Journal of Digital Information*, Volume 1 Issue 8, 2.

Walker, D.E., 1987. Knowledge resource tools for accessing large text files. In Sergei Nirenburg (ed.), Machine Translation: Theoretical and methodological issues. pp.247-261. *Cambridge: Cambridge University Press*

Wiener, N.,1961. Cybernetics, or Control and Communication in the Animal and the Machine. New York and London: *M.I.T. Press and John Wiley and Sons, Inc*.

Wolfram, D., Zhang, J., 2008. The Influence of Indexing Practices and Weighting Algorithms on Document Spaces. Journal of The American Society for Information Science and Technology, 59(1):3–11.

APPENDIX A

Medline abstract ID = 21116432.

Title: Expert public health nursing practice: a complex tapestry.

Abstract: The research outlined in this paper used Heideggerian phenomenology, as interpreted and utilised by Benner (1984) to examine the phenomenon of expert public health nursing practice within a New Zealand community health setting. Narrative interviews were conducted with eight identified expert practitioners who are currently practising in this speciality area. Data analysis led to the identification and description of themes which were supported by paradigm cases and exemplars. Four key themes were identified which captured the essence of the phenomenon of expert public health nursing practice as this was revealed in the practice of the research participants. The themes describe the finely tuned recognition and assessment skills demonstrated by these nurses; their ability to form, sustain and close relationships with clients over time; the skillful coaching undertaken with clients; and the way in which they coped with the dark side of their work with integrity and courage. It was recognised that neither the themes nor the various threads described within each theme exist in isolation from each other. Each theme is closely interrelated with others, and integrated into the complex tapestry of expert public health nursing practice that emerged in this study. Although the research findings supported much of what is reported in other published studies that have explored both expert and public health nursing practice, differences were apparent. This suggests that nurses should be cautious about using models or concepts developed in contexts that are often vastly different to the New Zealand nursing scene, without carefully evaluating their relevance.

APPENDIX B

"First" Theorem on Ontology. If set S_1 is semantically related to S_2, then for each cover $\mathrm{SSC}(S_1)$ and $\mathrm{SSC}(S_2)$, $\mathrm{SSC}(S_1) \cap \mathrm{SSC}(S_2) = \emptyset$, there are $T_1 \in \mathrm{SSC}(S_1)$ and $T_2 \in \mathrm{SSC}(S_2)$ that are semantically related.

Proof. Let L be the path of links from $T_a \in S_1$ to $T_b \in S_2$ that are in $SSC(S_1 \cup S_2)$ and T be the last node from $SSC(S_1)$ before next node on L is not from $SSC(S_1)$ when moving from T_a to T_b. We reassign $T_a = T$. Let next node on L after T_a be a child for T_a; the opposite case when next node is a parent for T_a is considered analogously. If next node for T_a is T_b then T_a and T_b are semantically related and the proof is complete. Otherwise, let T_c be the last child for T_a on L before next node on L is a

parent for T_c. By construction, $T_c \in SSC(S_1 \cup S_2)$ thus there is a child $T \in S_1 \cup S_2$ for T_c. Again, if $T \in S_2$, the proof is complete, else we reassign $T_a = T$ having $T_a \in S_1$, T_a is a child for T_c, and next node on L after T_c is a parent for T_c. Now let T be the last parent for T_c on L before next node is a child for T that is not in S_2. $T \in SSC(S_1 \cup S_2)$ and we will repeat our argument until after finite number of steps we either complete the proof, or reach $T_b \in S_2$ which is a child of a parent T on L that is also a parent of $T_a \in S_1$ by previous constructions, and this finally completes the proof.