Measuring Relatedness and Augmentation of Information of Interest within Free Text Law Enforcement Documents

Johnson, James R. (Bob); Miller, Anita
ADB Consulting
Carson City, Nevada, USA
james_r_johnson@earthlink.net

Khan, Latifur; Thuraisingham, Bhavani
University of Texas at Dallas
Richardson, Texas, USA
lkhan@utdallas.edu

***Abstract* -- This paper defines and shows the merit of measures for quantifying the degree of relatedness of information of interest and the importance of new information found within a large number of free text documents. These measures are used for identifying and sorting free text documents that are found to contain related information of interest and, in some cases, new information of interest related to a reference document. The relatedness measures consider the semantic content (e.g., people, vehicles, events, organizations, objects, and locations with their descriptive attributes) as well as the semantic context between semantic content items and key entities such as events and temporal items. Additional links to related sub-graphs between a reference graph and a comparison graph identify augmented knowledge over the known semantic text. Graph structures are generated initially from syntactic links and ontological class hierarchies, and augmented by inferred links resulting from triggered DL-Safe rules and abductive hypotheses. Inferred context broadens the potential for detecting related information. The approach is tested on a large set of free text emails between law enforcement detectives seeking leads for solving cases but the research has broad applicability to other domains such as intelligence collection, investigative reporting, and media monitoring.**

Keywords – relatedness measure; semantic content; semantic context; semantic sub-graphs; natural language processing, graph matching; free text; law enforcement; ontology; information of interest; abductive reasoning; semantic information structure; augmented knowledge.

I. Introduction

This paper describes the culmination of a three-year project addressing the identification of related information of interest across free text documents. Five research papers, [1], [2], [3], [4] and [5], address foundational research leading to this paper. This paper brings these research components together to provide three relatedness measures for identifying common pieces of information as well as new information augmenting the current knowledge base.

Related research is described in the next section. The class of problems being addressed; the test data used; and the objectives of the research are described in the following sections. An overview of the foundational functionality and terminology provided by previous project research is also included.

A. Related Research on Relatedness Measures

Measures of relatedness have been investigated previously. The focus has typically been on document-level categorization for the purpose of query expansion rather than on the focus of this research – the identification of related information of interest within documents of any category. Examples of previous research include the interestingness measures defined in [6] and [7] for ranking discovered knowledge. Interestingness measures typically utilize word frequencies and probability distributions (used in heuristic measures) and not domain-specific relationships needed in semantic measures. Reference [8] explored the performance of many semantic relatedness measures and found that their performances were inconsistent. Examples of relatedness measures employed for rating relatedness between documents include Latent Semantic Analysis [9] which relies on the tendency of words to appear in the documents being evaluated; Wikipedia-based measures [10] which are similar to thesauri-based approaches; Explicit Semantic Analysis [11] which compares weighted vectors from Wikipedia articles to each term; and the use of the quantity of Wikipedia hyperlink structures to define a relatedness measure [12]. Although Wikipedia spans a very large and growing topical list, specific domain terminology, acronyms, and abbreviations, such as employed in the law enforcement domain, are not present. Wikipedia sources are often used to provide semantic links for extracted entities within a document [13]. These provide links to external free text documents but do not integrate semantic context links.

B. Class of Problems Being Addressed

This research addresses situations involving large numbers of unstructured (free text) documents where the documents containing related pieces of information of interest as well as augmented knowledge are sought. Potential applications include detecting conversations and comments on a subject of interest to a researcher in social media; identifying

This work was sponsored by the Air Force Office of Scientific Research under Contract FA-9550-09-0468.

978-0-7695-4782-4/12 $26.00 © 2012 IEEE
DOI 10.1109/EISIC.2012.49

intelligence reports that include information on a situation of interest; and finding news articles that include information on an item of interest to a writer. These measures can also be used by web search engines to build semantically-based, contextually-related or inference-driven indexes. This application is different from document categorizing [14] where word frequencies are employed. Categorizing fails to identify pieces of information within documents of different categories. This application is also different from clustering where sentiment analysis is applied.

C. Test Data Used

The data used in this research was purposefully selected because it exhibits the full spectrum of complexities encountered in this class of problems. The test data is comprised of law enforcement emails exchanged between law enforcement investigators across a large geographical region. The investigators are typically sending out selected information on their current cases and they are seeking critical pieces of related information and leads for solving cases from historical emails. The complexities exhibited include: (1) related information of interest may be located within documents that are on unrelated subjects and that are very different types of documents; (2) a general lack of punctuation and a lack of consistent grammatical constructs; (3) liberal use of capitalization, domain-specific terminology, acronyms, abbreviations, and slang; (4) cut-and-paste text insertions; and (5) a wide and unpredictable range of terminology used to describe information of interest such that a simple key word search is extremely ineffective. The complexities of the emails can result in processing issues such as: entity misnaming due to misspellings, attribute mixing from adjacent entities, confusion from unrelated words surrounding entities, syntactical link extraction errors, false triggering of ontological rules, and incorrect assignment of semantic content and context items. These issues can translate to clutter in the subsequent graphs and hence uncertainties associated with the final relatedness measures.

D. Objectives of the Current Research

The objectives of the research described in this paper are to (1) detect related information of domain interest, (2) augment the knowledge base, in this case, linking to new leads for an investigator, and (3) sort the compared documents based on the degree of relatedness of the information of interest that they contain as well as the importance of any additional information found, in this case, giving higher priority to leads that match most closely or that augment the investigator's knowledge base the most.

E. Functionality Provided by Previous Research

An overview of the foundational research provided by the team's previous research is given in this section.

1) Identification of entities and their associated attributes (semantic content): Semantic content was defined in [2] as the union of an entity phrase and associated attributes, and is represented using an expanded entity phrase structure, namely,

$$\textit{Semantic Content Item} = \{EPT{:}c,\ W_a,\ P_1{:}A_1,\ \ldots,\ P_n{:}A_n\} \qquad (1)$$

where EPT = entity phrase tied to an ontology class, W_a = anchor word(s) in the text, if they exist, and A_i = i^{th} attribute (value) associated with the i^{th} property P_i.

Candidate entities are identified as a result of regular expression and lookup table data applied to part of speech tagging [2]. An ontology consisting of a hierarctical structure of concepts connected to a detailed Thesaurus aids in entity categorization [3]. Detected boundary terms and search windows are used for entity phrase growth and attribute assignments. Early analysis showed that the comparison of entities alone was ineffective but when the associated attributes were considered as well, the information was much more valuable. A simple example would be the difference between finding the word "man" as compared to also finding that the man had red hair and a distinctive tattoo on his neck.

2) Selection of semantic content that is of interest: Reference [3] explains that semantic content items are retained if they are of interest as specified by the ontology and the associated Thesaurus. Which information is classified as "information of interest" is domain-specific. In the case of the law enforcement investigators emails, for example, investigators identified their key interest areas as people, vehicles, events, organizations, objects, and locations. Other entity types are needed for identifying entities such as email addresses, phone numbers, group names, drug types, and business names. A semantic content item is determined to be "of interest" if the entity corresponds to an ontology subclass (handgun subclass of weapon class, for example) or instance (Glock instance of handgun, for example).

3) Handling of slang, acronyms, and domain-specific terminology: Many data issues were addressed in [3] by the addition of an ontology comprised of existing domain-independent ontologies and of new domain-dependent ontologies populated manually through law enforcement materials and reviews. A domain-specific Thesaurus linked to the ontology includes slang, acronyms, and domain-specific terminology. For example, an acronym such as BMV (for burglary of a motorized vehicle) is linked to a subclass of criminalEvent.

4) Inference through rules and hypotheses: Inference can expand the information and improve the check for relatedness. Inference is accomplished through mechanisms for capturing domain-specific processes: hierarchical inheritance, Description Logic (DL) Safe rules defined on an ontology, and abductive hypotheses with observation templates. Examples of inference in the law enforcement domain include:

- inheritance of person attributes by a suspect subclass,
- connection of a person to an event using syntactically connecting words, and
- identification of a person as a suspect via a relation in a DL-Safe rule.

5) Quantification of the importance of information found: Reference [4] developed an importance measure for semantic content which is used to facilitate quantitative comparison and prioritization between semantic content items. The importance measure is domain-specific and assigns more importance to attributes that contribute to unique identification of the associated entity. Specifically, the importance factor is computed as the inverse of the number of potential values. To illustrate, attributes that uniquely identify a person (eg. drivers license number) are more important than general attributes such as hair color.

6) Identification of relationships between entities using semantic context: The addition of contextual information in [4] was an important step to improve the prioritization of results and identification of new information. Entities with attributes (semantic content item) are treated as objects. However links between these entity-based objects incorporate their context relative to the global text structure. A document may contain more than one set of connected links of entities. In this research an entty with a large number of outgoing links is used as a starting point for collecting all connected entities into a structure called a semantic context item. This provides a basis for linking entities in context. Section E.8 organizes these semantic context items so they can be compared across documents.

A semantic context item is defined as:

A semantic context item is the set of all links and secondary links from the entity with the maximum residual-degree where residual-degree, r(i) is

$$r(i) = R^{out}(i) - R^{in}(i) \tag{2}$$

Borrowing some terms from graph theory, the number of links outgoing from an entity, i, is called its out-degree, $R^{\mathbf{out}}(i)$. Likewise the total number of incoming links from an entity is called its in-degree, $R^{\mathbf{in}}(i)$. Their difference is the residual-degree of an entity.

If the residual-degrees of entities are sorted from maximum to minimum, then the entity with the maximum residual-degree is selected as the initial semantic context entity. The same process is repeated after all links to the initial semantic context entity are removed from the link list. In this way, disjoint sub-graphs are successively constructed.

Semantic context links between semantic content items may provide valuable information. For example, a person with certain attributes may be observed with a particular vehicle, where the person content item may not contain sufficient attributes to be of interest, but when linked with the vehicle, the person may become of great interest to a detective.

In [15] and [16], semantic content and semantic context were not treated separately. For example, [16] defines a semantic graph as having node, link and direction where semantic cases specify the link label types. These link label types are based on the discourse, the genre, the author, or even the particular situation. In [16], the focal node of the semantic graph is generally a node corresponding to a process and typically does not link to an ontology class. The research described in this paper separates these two groups to aid in the identification of related information (typically in the form of semantic content), the identification of new information, and prioritization of results using the semantic context items.

7) Quantification of the significance of semantic context relationships: A significance measure for a document is intended to aid in prioritizing related information of interest with respect to the specific domain. The sum of all linked semantic content relatedness measures and their respective link functions is selected as the significance measure as developed in [4].

8) Graphical representation of semantic information structures so graph theory approaches can be used to compare semantic information: Finding related and new information between a reference and comparison document reduces to identifying related semantic content items and semantic context links between the two documents. These constructed links can be represented as graphs [5]. Reference [5] develops the mathematical foundation for the representation of the semantic information structure as a graph. The definition of a graph was expanded in [5] to include the semantic relations between the nodes, namely, (1) ontological class hierarchies providing inheritance, (2) DL-Safe rules, (3) abductive hypotheses, and (4) syntactic patterns. The definition of a graph incorporating these semantic structures is

$$R = (V_R, W_R, E_R) \tag{3}$$

where V_R = {u: u is a node}, W_R = {v: v is an edge function or label}, and E_R = {e: e ∈ $V_R \times W_R \times V_R$}. The relatedness measure problem is addressed using graph matching where the relatedness measure is a function on the set of ordered pairs $E_R \times E_C$. E_R is the set of edges from the reference graph R constructed from the reference document, and similarly for E_C of the comparison document. Explicitly, the graph matching problem is stated as the following:

Find all disjoint matched sub-graphs and their relatedness measures $g(E_R \times E_C) \rightarrow \boldsymbol{R}^n$ on the Cartesian product, $E_R \times E_C$. (4)

The reason for finding all disjoint matches is that there can be more than one semantic context item (connected graph) within a document. The graph representation of the semantic information reflects these disjoint sub-graphs. Defining the range of the relatedness function to be $\mathbf{R^n}$**,** the quantification of the importance measures associated with the various semantic content classes such as people, vehicles, events, objects, and locations. Mapping the relatedness onto an n-dimensional space allows sorting by a category of most interest to the user.

II. Relatedness Measures

This section describes two relatedness measures (semantic content relatedness and semantic context relatedness of links between two documents) which are used to compute the composite relatedness. This section also describes a common sub-graph characterization (common links) of the semantic information structures that optimizes the sorting of comparisons across many documents. The graph theory terminology “node” and “edge” are used in this section, rather than the natural language processing terms “entity” and “link”, to adapt to the mathematical terminology of graph theory.

Note the following graph theory generalizations that apply to both the semantic content relatedness measure and the semantic context relatedness measure:

- The inputs to the relatedness measures are the semantic information structures extracted from the documents and described in [4]. Each semantic information structure is a data structure that contains the semantic content items with their associated measures of importance and the semantic context items with their associated measures of significance for each contextually linked item.
- The semantic information structure is converted to a mathematical graph format that includes numbering of the nodes, edges, and sub-graphs in preparation for graph theory-oriented operations.
- The resulting graph nodes and edges are compared pair-wise [5]. For example, to determine the graph most related to graph 1, graph 1 is compared with graph 2, then graph 1 is compared with graph 3, etc. The comparison results are collected in two matrices, a node-node matrix and an edge-edge matrix.

Fig. 1 shows pictorial representations of mathematical graphs for two example emails. The nodes are shown in rectangles with their attributes identified beside the rectangle. The edges are shown with labeled directed lines (e1, e2, …). Fig. 1 is used in this section to illustrate the measures.

A. Semantic Content Relatedness Measure

The semantic content relatedness measure between documents is the comparison of semantic content items such as people, vehicles, events, organizations, objects, and locations along with their associated attributes [5]. The semantic content relatedness process does pair-wise comparisons of the associated attributes of nodes in the reference graph and the comparison graph and also considers their importance. A node relatedness matrix is constructed where each matrix element is the relatedness measure between the i^{th} node in the reference graph to the j^{th} node in the comparison graph. Thus a matrix element a_{ij} is

$$a_{ij} = \sum_k \sum_l \left(x_{ik} s_{kl} y_{jl} \right) \qquad (5)$$

where x_{ik} is the importance of the k^{th} attribute of the i^{th} node from the reference graph, s_{kl} is the similarity (exact, Levenshtein, synonym match (using lookup table from Thesaurus) between the attribute values from each node, and y_{jl} is the importance of the l^{th} attribute of the j^{th} node from the comparison graph. The Thesaurus includes abbreviations, acronyms, and slang. Note that the matrix elements are zero if the two semantic content items are not of the same type.

The elements of the semantic content matrices for graphs 1 and 2 in Fig. 1 are computed using (5). For example, if we assume that the importance value for color is 4 and the importance value of the driver’s license is 100, then the matrix element (1,1) is 4*1*4 + 100*1*100 + 1 + 1 + 1 = 10019, where the similarity between actual words is an exact match.

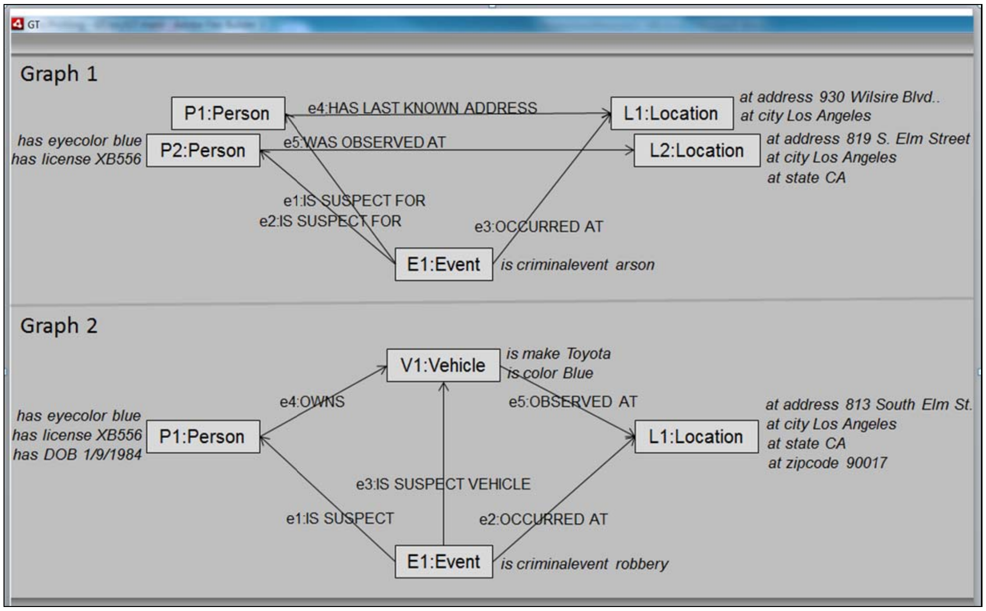


Figure 1. The use case comprised of graph 1 (upper) and graph 2 (lower) shows the semantic content and semantic context details of two free text documents.

The semantic content relatedness matrix in Fig. 2 shows strong relatedness measures for P2:Person (graph 1 representation in Fig. 1) to P1:Person (graph 2) since the license numbers match exactly. Much less relatedness is shown between node 4 (graph 1) and node 3 (Graph 2) where a portion of the location attributes match. Since comparisons never yield matches when the entities are of different types, the resulting matrix values are 0.

Semantic content relatedness matrix elements that are nonzero can belong to one of the class types. The relatedness measures can be mapped into $\boldsymbol{R}^n$ where i=1...n is the class dimension. A vector comprised of the sum of all matrix elements from each type (dimension) represents the total relatedness measure for each type. Therefore the vector A has components

$$A_k = \sum_{l=1}^{n} (a_k)_l \qquad (6)$$

where k denotes the class type and l sums over all the elements of that type. The length of **A** is then the total semantic content measure across all types, namely,

$$M_{Content} = \sqrt{\sum_{k=1}^{n} (A_k)^2} \qquad (7)$$

Using the data in Fig. 2, the total semantic content relatedness measure is 10,028.8 with the (1,1) element totally overwhelming this value. Note that the measure includes all nodes regardless of whether or not they are part of the same sub-graph. Also note that the measure can be used as or as part of a sort parameter for sorting documents with related information of interest.

B. Semantic Context RelatednessMeasure

The semantic context relatedness measure is important in that it identifies common contextual relations between two documents. The objective of the common semantic context relatedness measure is to determine whether there are common edges between related nodes in both graphs. When this situation is found, it indicates relatedness between links of content items. For example, a suspect linked to a certain vehicle may appear in multiple documents.

A semantic context relatedness matrix was introduced in [5] where the elements are the relatedness between corresponding edges within each graph. Note that only edges which have related nodes and edge functions have nonzero matrix elements. In mathematical form these matrix elements are computed from the following:

$$p_{ij} = g_{ij} \cdot m(f_i) M_{ij} m(f_j) \cdot \left(\sum_{k=s,t} \sum_{l=s,t} (y_k r_{kl} y_l) + \sum_{n=1,N} \sum_{m=1,M} (x_{kn} s_{nm} x_{tm}) \right) \qquad (8)$$

where s_{kl} is the similarity measure between two attribute labels; r_{kl} is the similarity between node labels; and M_{ij} is the similarity between edge labels. The factors y_k denote the level of semantic content measure associated with the kth node. The factor x_{kn} denotes the importance associated with the nth attribute of the kth node. E_k denotes a node label with S_{kl} being the similarity between two node labels. The contribution of the significance is factored by g_{ij}. The semantic context relatedness matrix corresponding to the graphs in Fig. 1 is shown in Fig. 3.

$$\begin{array}{cc|cccc} & & \multicolumn{4}{c}{\text{Graph 2 Nodes}} \\ & & 1 & 2 & 3 & 4 \\ \hline \text{Graph 1 Nodes} & 1 & 10019 & 0 & 0 & 0 \\ & 2 & 0 & 2 & 0 & 0 \\ & 3 & 0 & 0 & 27 & 0 \\ & 4 & 0 & 0 & 393 & 0 \\ & 5 & 1 & 0 & 0 & 0 \end{array}$$

Graph 1 Nodes:
1 P2:Person
2 E1:Event
3 L1:Location
4 L2:Location
5 P1:Person

Graph 2 Nodes:
1 P1:Person
2 E1:Event
3 L1:Location
4 V1:Motorized Vehicle

Figure 2. Semantic content relatedness matrix from graphs in Fig. 1

Edge 1 in graph 1 (E1:Event -> P2:Person) has a very high relatedness to edge 1 of graph 2 (E1:Event -> P1:Person). The high match corresponding to the (1,1) matrix element in [3] is due to the match of the attributes of the person (license) as well as the match of the link function. The semantic context relatedness matrix identifies the degree of relatedness on an edge-by-edge basis. In Fig. 3, the rows correspond to the labeled semantic context links in graph 1 (Fig. 1), while the columns correspond to the labeled semantic context links in graph 2 (Fig. 1). The matrix element values are computed from (9). The significance factor is determined by a function of the importance factors and the link function type [4]. The edge labels (term 1) in the (1,1) matrix element match exactly. With $m(f_i)$ = 1, this first term is 1. The second term is a measure of the node relatedness from the semantic content relatedness matrix (Fig. 2). For the (1,1) element this term measures the relatedness between the two E1:Event nodes, with a value of 1 for matching the Event spelling. The attributes are not equal and therefore do not contribute to this term. The last term measures the relatedness between the P2:Person (graph 1) and the P1:Person (graph 2) with a value of 10,019 (Fig. 2). Thus the (1,1) entry is 811,539.

$$\begin{array}{cc|ccccc} & & \multicolumn{5}{c}{\text{Graph 2 Edges}} \\ & & 1 & 2 & 3 & 4 & 5 \\ \hline \text{Graph 1 Edges} & 1 & 811539 & 0 & 0 & 0 & 0 \\ & 2 & 0 & 0 & 0 & 0 & 0 \\ & 3 & 0 & 27 & 0 & 0 & 0 \\ & 4 & 0 & 0 & 0 & 0 & 0 \\ & 5 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Figure 3. Semantic context relatedness matrix from graphs in Fig. 1

The semantic context relatedness matrix elements comprise measures from more than one class type. Hence a vector approach similar to the semantic content relatedness measure is not possible. For simplicity, the absolute value of the semantic context relatedness matrix elements summed resulting in

$$M_{Context} = \frac{1}{mn}\sum_i \sum_j |p_{ij}| \tag{9}$$

where m and n are the number of rows and columns, respectively. The absolute value removes issues concerning the edge directions. The coefficient 1/mn serves as a normalization factor across various document pair comparisons. In the example from Fig. 3, $M_{Context}$ = 811,564. Note that the measure can be used as or as part of a sort parameter to place emphasis on contextually linked information.

C. Common Sub-graph Characterization

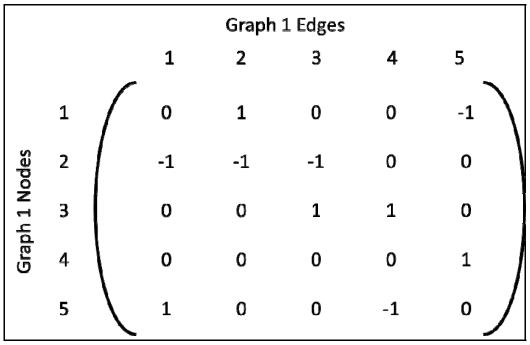
This section describes an approach for determining related sub-graphs between the graphs extracted from two documents. The objectives of the semantic common sub-graph characterization are two-fold: (1) support sorting of relatedness results by providing documents with most potential first and (2) enable detection of new information from connections not contained in the reference document but found to be connected in the comparison document. The semantic context relatedness matrix only determines if there are common edges between related nodes in two graphs. Note that the semantic sub-graph relatedness algorithm finds all disjoint related sub-graphs between the reference graph and the comparison graph through a graph matching technique. The disjoint related sub-graphs are obtained by combining the information in the semantic context relatedness matrix (related links from the two graphs) and the adjacency matrices which capture the starting and ending nodes as well as their directions.

The adjacency matrix $[A]_{ij}$ is defined with each row corresponding to an element of V_R and each column corresponding to an element of E_R such that the starting and ending nodes within a graph are identified.

$$[A]_{ij} = \begin{cases} -1 \;\; if \;\; s_A(j) = i \\ 1 \;\; if \;\; t_A(j) = i \\ 0 \;\; otherwise \end{cases} \tag{10}$$

where $s_A(i)$ is the source of the ith edge, and $t_A(i)$ is the terminus of the ith edge. The adjacency matrix for the first example graph in Fig. 1 is shown in Figure 4.

Note that each adjacency matrix has a -1 corresponding to the source (starting) node and a +1 corresponding to the terminal (ending) node of an edge. In graph 1, edge 1 (column 1) starts at node 2 (E1:Event) and ends at node 1 (P2:Person).



Graph 1 Edges

Graph 1 Nodes

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 1 & 0 & 0 & -1 \\ 2 & -1 & -1 & -1 & 0 & 0 \\ 3 & 0 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 \\ 5 & 1 & 0 & 0 & -1 & 0 \end{array}$$

Figure 4. Adjacency matrix for graph 1 in Fig. 1

An adjacency matrix provides the linkage description of each graph. The semantic relatedness matrix provides the information on which links are semantically related. These matrices allow construction of the related disjoint sub-graphs through the method described here.

The sub-graphs being compared generally have different numbers of nodes and edges. Not all nonzero elements of the semantic context relatedness matrix are highly related. By selecting the maximum matrix element in each column, partially related edges are eliminated in favor the most related edges. The method of finding existing common sub-graphs $G_1 \subset H_1$ and $G_2 \subset H_2$ is therefore the following:

> *Let $[S]_{ij}$ be the semantic context relatedness matrix. For each column, find the maximum value, if it exists. If more than one maximum exists, then no maximum is designated. Each maximum corresponds to related edges in each graph H_1 and H_2. Locate the corresponding edge in for each adjacency matrix ($[A_1]_{ij}$ and $[A_2]_{ij}$) resulting in a subset of edges in each sub-graph. Construct sub-graphs G_1 and G_2. Create $H_1 \backslash G_1$ (element in H_1 but not in G_1) and $H_2 \backslash G_2$ then repeat the process to extract other common sub-graphs, if they exist.* (11)

As an example, the maximum values in each column of the semantic context relatedness matrix are listed in Table I.

The maximum relatedness measures listed in Table I represent the connected pieces of information that are related between the two documents. The common sub-graphs corresponding to the related edges in Table II consist of edges e1 -> e1 and e3 -> e2, respectively in graphs 1 and 2 (Fig. 1).

TABLE I. MAXIMUM RELATEDNESS MEASURES IN THE SEMANTIC RELATEDNESS MATRIX FROM FIG. 3

Column	Max Location	Value
1	(1,1)	811,539
2	(3,2)	27
3	n/a	0
4	n/a	0
5	n/a	0

D. *Augmented Knowledge Measure*

The common sub-graphs G_1 (from reference document) and G_2 (from comparison document) may have additional connected edges which are not part of the common sub-graph matching. The additional edges connected to G_2 represent additional knowledge previously unknown through the reference document. The augmented graph is denoted by T_2 then referring to graph 2 in Fig. 1, the extended edges are listed in Table II.

The augmented knowledge measure is dependent on both the connectedness (multiple edges connected to the common sub-graph) and the significance of the new connections. The augmented knowledge measure, denoted by $M_{Augmented}$ (12). In the example, V1:Vehicle is connected to the common sub-graph by three edges. The expanded knowledge measure is the defined as the product of the number, N, of connecting edges and the sum, S, of the significance quantities for each new edge. Hence,

$$M_{Augmented} = N \cdot S. \tag{12}$$

For the example, $M_{Augmented} = 3 \text{ x } 18 = 54$. This value is quite large due to the constructed example in Fig. 1. Typically these measures are closer to 1 (see Fig. 6). Note that the measure can be used as or as part of a sort parameter for sorting documents with related information of interest to provide potentially valuable new knowledge first.

III. Evaluation Results

The research software to test the relatedness measures is over 17,000 lines of code. The development/research environment was written in Perl. Adobe Flex software was developed to support the visual assessment of the ground-truth and test result evaluation.

The test data, sanitized emails from a large LISTSERV network connecting nearly 1000 local, state, and federal law enforcement agencies covering over 4 years, is further described in Section B of the Introduction. The sanitization involved manually constructing a lookup table changing names, addresses (to other actual addresses), phone numbers and other identification items.

763 email documents were processed. The average document size was 144.7 words with an average of 56.1 entities extracted per document. The software was able to name 70% of the entities as a person, vehicle, event, organization, object, or location. The average number of related nodes per file pair was 25.4. Table III provides statistics characterizing the data extracted.

TABLE II. Augmented Graph 2 Edges from Fig. 1

Edge Number	Edge Description
e3	E1:Event -> IS SUSPECT VEHICLE -> V1:Vehicle
e4	P1:Person -> OWNS -> V1:Vehicle
e5	V1:Vehicle -> OBSERVED AT -> L1:Location

TABLE III. Statistics Characterizing the Data Extracted

Statistic	Value
Total number of words	115,011
Total number of entities extracted	44,620
Number of entity types extracted	17
Number of persons extracted	12,544
Number of vehicles extracted	3,217
Number of events extracted	702
Number of organizations extracted	2284
Number of locations extracted	1263
Number of links extracted	7,959
Number of related content items detected	27,039

The data was processed so that each file's semantic information structures were compared to that of the remaining files. 315,615 file comparisons were made. The relatedness measures divided into three groups: (1) those with no relatedness, (2) those with low relatedness corresponding to no real connections, and (3) those with high relatedness where significant connections between semantic content and context occurred. For this law enforcement data, the threshold for separating groups 2 and 3 is approximately 4000. Fig. 5 shows the histogram of graph node matches, where the x-axis is the total relatedness measure; the y-axis is log number of matches; groups 1 and 2 are indicated to left and right of vertical dashed line, respectively.

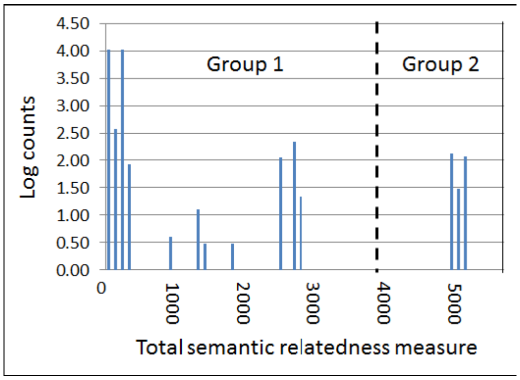


Figure 5. Total semantic relatedness measure

Fig. 6 shows the histogram of the augmented knowledge measure corresponding to file pairs having highly related semantic content and semantic context. This histogram shows that over 50 percent of the data in Group 2 contributes augmented knowledge.

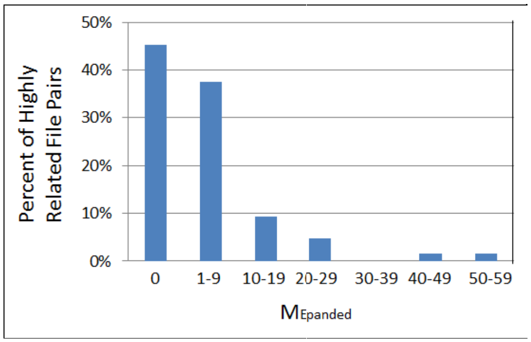


Figure 6. Histogram of augmented knowledge measure

Examples of the valuable additional information discovered include the following:

- A case where an investigator was seeking burglaries of a motor vehicle that might be related to his investigation. Seven related events in his geographic area were detected. One result matched the crime and location, and provided additional information of interest on a vehicle that was involved, the business parking lot, and the date of the event.
- In many cases, emails containing valuable related information of interest were found, even when the criminal event was not the same and no person was named.
- Emails were identified describing events where a similar vehicle was involved but different suspects were identified, providing the investigator new leads to follow.
- Additional information was often identified as part of new knowledge. For example, a gun used in an aggravated assault, was identified in another email which also provided the suspect's last known address, the complainant's name and information about the money involved.

The sorting algorithm for results may vary depending on the application. In the case of investigators seeking new leads, the sort algorithm selected was to sort the historical documents with information of interest based on the semantic content relatedness measure first, then the augmented knowledge measure, and then the semantic context relatedness measure.

IV. CONCLUDING REMARKS

The results of this approach demonstrate that use of semantic relatedness measures and a measure of augmented information of interest is very effective at identifying and prioritizing semantically significant information for the user. The key areas of success of this approach are:

- clean extraction of named entities and their attributes, and identification of links between entities
- generation of multi-connected nodes within the graph that provide higher potential for augmented knowledge

This research has also spawned many interesting derivative research areas such as:

- application of graph matching to identify topical trends and conversational flow,
- learning new classes and relations for ontology growth,
- generalization of ontologies and thesauri to minimize domain-specific content,
- application to mixed sources (eg. Email, news feeds, and formal reports),
- link extractions crossing complex and compound sentence boundaries, and
- application to languages other than English.

REFERENCES

[1] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, and M. Kantarcioglu, "Identification of related information of interest across free text documents," IEEE Intelligence and Security Informatics, Beijing, July 2011.

[2] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, and M. Kantarcioglu, "Extraction of expanded entity phrases," IEEE Intelligence and Security Informatics, Beijing, July 2011.

[3] J. Johnson, A. Miller, and L. Khan, "Law enforcement ontology for identification of related information of interest across free text documents", IEEE European Intelligence and Security Informatics, Athens, September 2011 .

[4] J. Johnson, A. Miller, L. Khan, and B. Thuraisingham, "Extracting semantic information structures from free text law enforcement data," IEEE Intelligence and Security Informatics, pp. 177-179, Washington, D.C., June 11-14, 2012.

[5] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, "Graphical Representation of Semantic Information," International Conference on Semantic Computing (submitted), Palermo, September 2012.

[6] R. J. Hilderman, and H. J. Hamilton, ``Heuristic Measures of Interestingness." In Third European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'99), Prague, Czech Republic, Springer Verlag, September, pp 232-241, 1999.

[7] B. Barber, and H. J. Hamilton, ``Extracting Share Frequent Itemsets with Infrequent Subsets," Data Mining and Knowledge Discovery, vol 7(2), pp153-185, April 2003.

[8] I. Cramer, "How Well do Semantic Relatedness Measures Perform? A Meta Study," In: Bos, J. / Delmonte, R. (eds). Semantics in Text Processing: STEP 2008 Conference Proceedings (Research in Computational Semantics). London: College Publications. 59-70.

[9] T. K. Landauer, P. W. Foltz. and D. Laham, "An introduction to latent semantic analysis," Discourse Processes, Vol 25(2-3), pp 259-284, 1998.

[10] M. Strube and S. P. Ponzetto, "WikiRelate! Computing Semantic Relatedness Using Wikipedia," In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), pp.1419-1424, 2006.

[11] E. Gabrilovich and S. Markovich, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.

[12] D. Milne and I. H. Witten, "An Effective, Low Cost Measure of Semantic Relatedness Otained from Wikipedia Links," In Proceedings of the First AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08), Chicago, Il, 2008.

[13] C. Giannone, R. Basili, P. Naggar and A. Moschitti, "Supervised Semantic Relation Mining from Linguistically Noisy Text Documents," International Journal on Document Analysis and Recognition (IJDAR) DOI 10.1007/s10032-010-0138-0.

[14] H. H. Tar and T. T. S. Nyaunt, "Ontology-based Concept Weighting for Text Documents," World Academy of Science, Engineering and Technology, Vol 81, 2011.

[15] J. Sowa, "Conceptual Graphs," Handbook of Knowledge Representation, ed. By F. van Harmelen, V. Lifschitz, and B. Porter, Elsevier, 2008, pp. 213-237.

[16] L.Hébert, "Dispositifs pour l'analyse des textes et des images," Limoges, Pulim, 2007, 282 pages, Publié en ligne le 5 décembre 2007.