An Ontological Approach to Assessing IC Need to Know

Phillip Burns CTA Inc.
Prof. Amit Sheth LSDIS Lab, University of Georgia
Presented to ARDA PI Meeting, Myrtle Beach, February 16 2005

Contract # NBCHC030083

A thought to begin with …

- You cannot separate two facets of information retrieval ("systematic serendipity")—information recovery and information discovery.
  - Eugene Garfield … in essays of an Information Scientist
Objective & Approach

- Determine if (classified) documents reviewed by an IC analyst satisfy his/her “need to know”
  - Characterization of “need to know” w.r.t. ontology
  - Characterizing document content in terms of ontology
  - Discovering weighted semantic relationships between document content and “need to know” characterization

Characterizing “Need to Know” using a Semantic Approach (using Ontology)

- Requires domain ontology
  - models important concepts & relationships of domain (schema), captures factual knowledge (instances)
- Relate analyst’s need to know to concepts & relationships in ontology
  - e.g. terrorist organization, funding sources, facilitators, members, methods
Characterizing document content in terms of ontology: “Semantic Annotation”

- Correlate words/phrases from document with concepts/relationships in ontology
- Meta-data added to document (from associated ontological knowledge)
- Active area of research but practically useful technology now available (e.g., Semagix Freedom)

Semantic Relationships between Document & “Need to Know”

- Semantic associations: relationships between document concepts & “need to know” concepts are discovered and ranked
- Ranking based on multiple factors
  - no. of links, types of links, location in ontology, …
- Ranking indicates degree of semantic “closeness”
  - and therefore, how related document is to “need to know”
Research Content

- Discovery & ranking of semantic associations
- Characterizing “need to know” in terms of ontological concepts & relationships (context of investigation)

While applying emerging technologies for
- Ontology design and population
- Meta-data annotation of heterogeneous documents
  - correlation of document content with concepts in ontology

Relevance Ranking of Documents

Four groups of document-ranking:
- **Not Related Documents**
  - unable to determine relation to context
- **Ambiguously Related Documents**
  - some relationship exists to the context
- **Closely Related Documents**
  - Entities are closely related to the context
- **Highly Related Documents**
  - Entities are a direct match to the context

Cut-off values determine grouping of documents w.r.t. relevance
- These are customizable cut-off values (more control and more meaningful parameters compared to say automatic classification or statistical approaches)

“Inspection” of a document is possible via (a) original document or (b) original document with *highlighted* entities
**Relevance Function (w.r.t. Context)**

"Closely related entities are more relevant than distant entities"

\[ E = \{ e \mid e \in \text{Document} \} \]

\[ E_k = \{ f \mid \text{distance}(f, e) = k \} \]

\[
\text{Document Relevance} = \sum_{i=0}^{\infty} \left( \begin{array}{c}
\text{classes Relevance}(E_i) \\
\text{relations Relevance}(E_i) \\
\text{entities Relevance}(E_i)
\end{array} \right)
\]

---

**IA Context of Investigation**

(characterization of “Need to Know”)

We define the context of investigation as a combination of the following:

- A set of entity classes and relationships, and/or a negation of a set of entity classes and relationships
- A set of entity instance names, and/or a negation of a set of entity instance names
- A set of keyword values that might appear at any attribute of the populated instance data, and/or a negation of a set of keyword values
Goal is to capture, at a high level, the types of entities, (or relationships), that are considered important.

Relationships can be constrained to be associated with specified class types

E.G. It can be specified that a relation ‘affiliated with’ is part of the context only when it is connected with an entity that belongs to a specific class, say, ‘Terror Organization’

---

graph-based creation of a context of investigation

26,489 entities
34,513 (explicit) relationships
Additional Semantic Constraints

Components of Document Relevance

1. Entities belong to classes in the Context
   \( \text{type(entity)} \subseteq \text{Context} \)

2. Relationships constrain
   \( \text{Relationship} \Rightarrow \text{[Class]} \)

3. Entities match a list of entities of interest (in the Context)
   \( \text{entity} \subseteq \text{Entities-List} \)
Some thoughts along the way

- “An object by itself is intensely uninteresting.”
  Grady Booch, Object Oriented Design with Applications, 1991

- I might as well join my better known colleagues:
  “Relationship is at the heart of semantics. Ontology is at the heart of the Semantic Web.”
Show me the stuff …

here you go … demonstration
Security and Terrorism Part of SWETO Ontology

Semantic Annotation

- Document searched for entity names (or synonyms) contained in ontology
- Then document entities are annotated with additional information from corresponding entities in ontology including named relationships to other entities
- Following chart is example
  - Highlighted text are entities found corresponding to concepts in ontology
  - XML is corresponding meta-data annotation
Relevance Measures for Documents
(relating document content to IA “need to know”)

- Relevance engine input
  - the set of semantically annotated documents
  - the context of investigation for the assignment
  - the ontology schema represented in RDFS, and
    the ontology instances represented in RDF

- Relevance measure function used to verify
  whether the entity annotations in
  the annotated document can be fit into
  the entity classes, entity instances, and/or
  keywords specified in the context of investigation.

Obama Bin Laden is both one of the CIA’s most wanted men and a hero to
his associates who are already being sought by the US on charges of interna-
tional year (1998) bombing of American embassies in Africa and last year’s
attack on the USS Cole in Yemen In May this year
a US jury convicted four men believed to be linked with Bin Laden of plotting the embassy bombings in Kenya and
Lampanga. Bin Laden, an increasingly wealthy and powerful man, has been granted a safe haven by Afghanistan’s ruling
Taliban movement. During his time in hiding, he has called for a holy war against the US, and for the killing of Ameri-
cans and Jews. He is reported to be able to rally around him up to 3,000 fighters. He is also suspected of helping to set up
Islamic training centres to prepare soldiers to fight in Chechnya and other parts of the former Soviet Union
Relevance Measures for Documents
(relating document content to IA “need to know” (cont))

- Documents classified as:
  - Highly relevant
    - Document entities directly related
  - Closely related
    - Document entities related through strong semantic associations
  - Ambiguous
    - Document entities related through weak semantic associations
  - Not relevant
    - Document entities not related to “need to know”
  - Undeterminable
    - Document entities not found in ontology

Challenges we have addressed

- Discovery of Semantic Associations per entity per document
- Input/Visualization/Management of Context of Investigation
- Scalability on number of documents & ontology size
  - Performs well (in terms of time and scalability) with thousands of documents and for scenarios when a IA investigation has involved hundreds of documents
- No systematic measure of quality for this specific application/scenario (general evaluation of research is done)
Challenges to be addressed

- Scalability to a million+ documents (possibly with preprocessing/filtering)
- Further development/enrichment of the ontology
- Improved measure of the strength of Semantic Associations
- Evaluations by human subjects
- Visualization and interactive discovery

References

7. C. Halaschek, B. Aleman-Meza, I.B. Arpinar, A. Sheth Discovering and Ranking Semantic Associations over a Large RDF Metabase Demonstration Paper, VLDB 2004, 30th International Conference on Very Large Data Bases, Toronto, Canada, 30 August - 3 September, 2004
References (cont)


Conclusions

- New Semantic Approach to a class of challenging problems: vendor vetting, knowledge discovery, ….
- Viability demonstrated on a small scale (comprehensive demonstration)
- Significant new research that builds upon the latest Semantic Platform
A parting thought

“Discovery commences with an awareness of anomaly …”
  - Thomas S. Kuhn,
    in The Structure of Scientific Revolutions