

Context based information trust analysis for threat detection

Vesile Evrim

DHS Center for Knowledge Integration and Discovery, University of Southern California

Introduction

Finding the most relevant and critical set of information in response to an information request of a Homeland Security analyst from unstructured open data in the availability of today's vast amount of digital data is a challenging problem.

- Documents can be about anything
- Anybody can be an information provider

Current Information Retrieval systems focus on satisfying the most users for most of their searches and leave aside the modeling of the context linked to the user's search. Thus, in this project, we use ontology-based approaches to analyze the domain information of the request, and user's interest context to present content and trust relevant results to the user.

Materials and methods

Many information request of an analyst are related to the events. Thus, in this study, information request of a user is analyzed in the context of events.

WH-Question	QA Event Elements
Who/Whose/Whom	Subject, Object
Where	Location
When	Time
What	Subject, Object, Description, Action

Table 1: Correspondence between selected WH-Questions & Event Elements

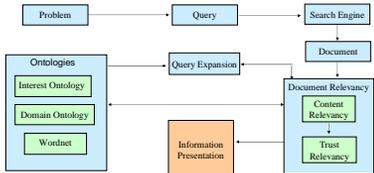


Figure 1: Context-Based Information Trust Analysis (CONITA) Framework

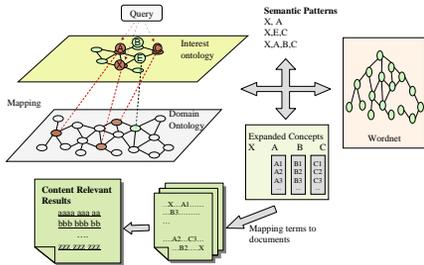


Figure 2: Steps in finding content relevant documents

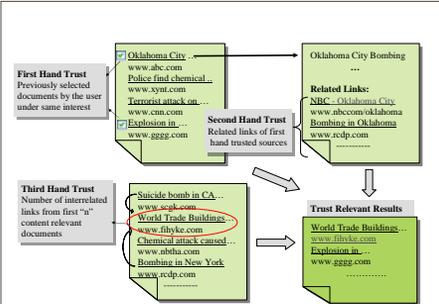


Figure 3: Trust relevancy based on first, second and third hand trust information for a given query in a certain interest context

Results

Overall 120 Homeland Security related queries are tested in Google, based on context dependency and the generality of the terms in the query (Table 2).

	General (22 queries)	Intermediate (12 queries)	Specific (6 queries)
Query Concepts	Threat Analysis	Threat Analysis	Threat Analysis
(event)	30%	33%	41%
(event, location)	37%	38%	45%
(event, location, time)	50%	47%	42%

Table 2: Relevancy of the query results for threat analysis based on specificity and contextually relatedness of the terms in the query

- Queries with more event context elements tend to return more relevant documents for threat analysis.
- As the specificity of the terms in the query increases the relevancy of the returned documents for threat analysis also increases.

Preliminary Case Study for Content Relevancy

Relevancy of the documents for user, search engine and CONITA is determined as follows:

Search Engine: First 10 documents returned by the search engine.
User: First 10 relevant documents selected by the user after analyzing the first "n" documents of the search engine's returned results.
CONITA: First 10 documents returned by CONITA after reprocessing the first "n" returned documents of the search engine.

Settings:

- 3 queries are tested on 2 different search engines and CONITA (Table 3).
- 5 users are trained about the domain to do the evaluations.

Problem	Find locations Laden involved in attacks
Query 1	<location, Laden, involved, attacks>
Problem 2	Find the individual terrorists involved in bombing attacks in Iraq
Query 2	<terrorist, involved, bombing, Iraq>
Problem 3	Find terrorist groups involved with biological attacks in America
Query 3	<terrorist group, involved; biological attack, America>

	Q1G	Google	CONITA	Q2G	Google	CONITA	Q3G	Google	CONITA
Precision	56%	42%	4%	6%	64%	82%			
	Q1Y	Yahoo	CONITA	Q2Y	Yahoo	CONITA	Q3Y	Yahoo	CONITA
Precision	70%	88%	28%	20%	12%	4%			

Table 3: Relevancy evaluation of the returned documents by CONITA, Google and Yahoo for the three given queries

Statistical results for 4 Queries		
Identical sources returned by Google, Yahoo (20 documents)		8%
Google - CONITA precision difference		17%
Yahoo - CONITA precision difference		20%
User evaluation's standard deviation		9%

Table 4: Statistical results of the study

On average, for the given queries, CONITA returns 20% more content relevant results for the information request of an intelligence analyst compared to other 2 information retrieval systems.

Preliminary Case Study for Trust Relevancy

In this study, it is assumed that user is a first time system user and no user history is available. Thus, the following results do not include first hand and dependent second hand trust information. CONITA's returned results for the following case, uses only third hand trust in calculating the trustworthiness value.

CONITA's trust relevancy is tested against the trust analysis of 5 evaluators over 15 documents. (Figure 4).

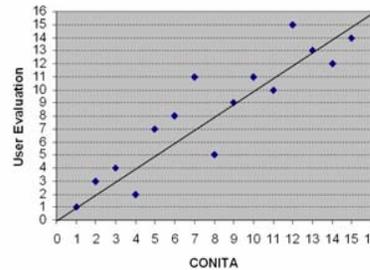


Figure 4: R-Square value for trust relevance between CONITA and user evaluation

- Standard deviation between 5 evaluators' results is 15%.
- The R-square value between the users' evaluation and CONITA is 0.81 which shows fairly close correlation.

Conclusions

Our preliminary results showed that context based information analysis increases the relevancy of the documents for the user's information request. Unlike most Information Retrieval systems, CONITA uses semantics behind the information request of an analyst to retrieve content and trust relevant results for the users.

Trust has a crucial role in an intelligence analyst's decision making process. However, because of the context dependency, trust has not been studied in detail by the Information Retrieval systems. CONITA uses the domain and user's context information to provide a customized trust measure for the users.

Currently CONITA is in the development phase and the results we provide here are preliminary. The success of the system is highly depend on the corpus (results returned by the search engine), ontologies used and the user feedback. We are in the process of utilizing a terrorism ontology to greatly increase relevance and look forward to testing the system with DHS analysts to better customize the framework for their needs.

Literature cited

Aleman-Meza, B., Burns, P., Evanson, M., Palaniswami, D. and Sheth, A. P. 2005. An Ontological Approach to the Document Access Problem of Insider Threat. IEEE International Conference on Intelligence and Security Informatics, Atlanta, Georgia, USA.

Allan, J. 2003. High Accuracy Retrieval from Documents. TREC: 24-37

Buneman, P., Khanna, S. and Tan, W. C. 2000. Data provenance: Some basic issues. In Foundations of Software Technology and Theoretical Computer Science.

DHS: Homeland Security. National Infrastructure Protection Plan. 2006.

Freund, L. and Toms, E. G. 2005. Contextual search: From information behavior to information retrieval. In Proceedings of the Annual Conference of the Canadian Association for Information Science.

Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J. 2004. Combating web spam with TrustRank. In proceedings of the 30th International Conference on Very Large Data Bases (VLDB), pp. 271-279, Toronto, Canada.

Hernandez, N., Mothe, J., Christmet, C. and Egret, D. 2007. Modeling context through domain ontologies. Information Retrieval 10, 2, 143-172.

Marsh, S. 1994. Formalizing Rust as a Computational Concept Ph.D. Thesis, University of Stirling.

Yang H., Chua T.S., Wang S. 2003. Modeling Web Knowledge for Answering Event-based Questions.

Acknowledgments

This project was funded through the Center for Knowledge Integration and Discovery by a contract from the Department of Homeland Security, Science and Technology Directorate, Office of University Programs.

For further information

Please contact evrim@usc.edu. More information on this and related projects can be obtained at Semantic Information Research Laboratory, <http://sir-lab.usc.edu>

