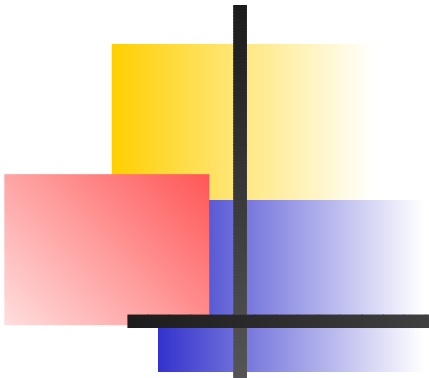


ClaimSpotter: an Environment to Support Sensemaking with Knowledge Triples

Bertrand Sereno, Simon Buckingham Shum & Enrico Motta

Knowledge Media Institute
The Open University
Milton Keynes MK7 6AA, UK

Claim spaces

Ontology-based debate and discussion: ScholOnto.

Project's goal: add a layer, sitting on top of a network of scholarly documents, composed of interpretations, personal notes and personal connections between scholarly documents.

These are expressed as semi-formalized statements (triples, also called **claims**), connecting 'concepts'.

Concepts are similar to tags. They are attached to a document.

Concepts (from a single document or from multiple ones) can be connected with a **relation** to form a claim. Relations are defined in an ontology of discourse.

Claim spaces

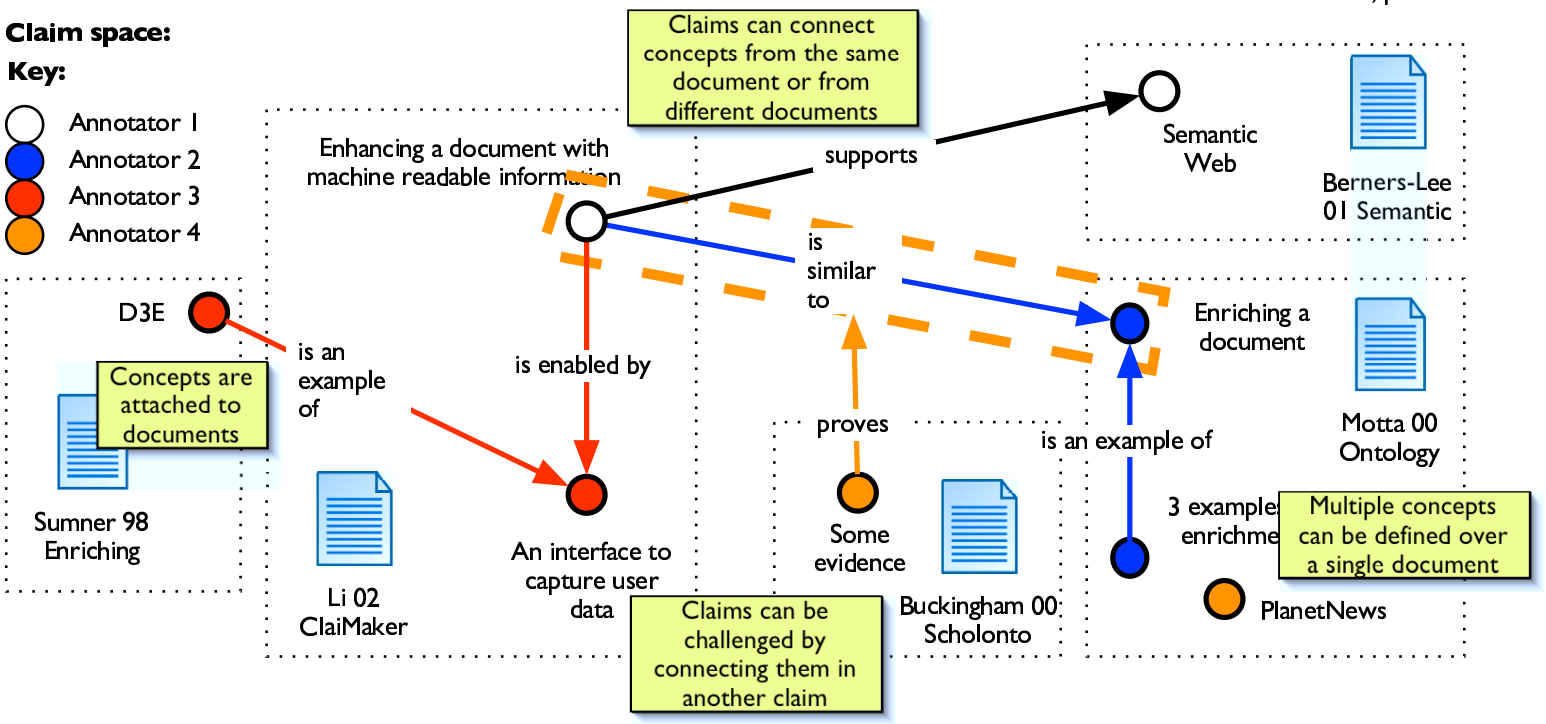
Concept types (optional):

analysis	methodology
approach	model
assumption	opinion
data	phenomenom
definition	problem
evidence	solution
hypothesis	theory
language	

Relation categories and types:

general: is about, uses/applies/is enabled by, improve on, impair
problem: address, solve
supports: prove, refute, is evidence for, is evidence against, agree with, disagree with, is consistent with, is inconsistent with
taxonomic: part of, example of, subclass of
similarity: is identical to, is similar to, is different to, is the opposite of, share issue with, has nothing to do with, is analogous to
causal: predict, envisage, cause, is capable of causing, is prerequisite of, is unlikely to affect, prevent

Claim space:



Claim spaces

Potential benefits:

- Follow the intellectual lineage of an idea
- Summarize the different approaches proposed to address a particular problem.
- Discover areas of agreement and debate

Costs:

- Interpreting a document is hard. Is translating one's opinion into a fixed set of triples going to be even harder ?
- Time. Effort.
- Who do I trust ?

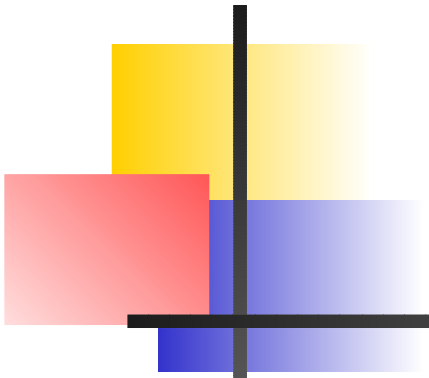
Claim spaces

What can we do to help the user bridge the gap between the richness of a scholarly document and a succinct set of ScholOnto claims ?

How can we support this translation process ?

Why is it difficult/different, compared to traditional document annotation projects ?

- The knowledge we are interested in capturing does not appear explicitly in the document but results from a sense-making process
- There is no truth, no 'correct' interpretation. To compare to approaches where a 'fact' has to be extracted from a document.
- Furthermore, this knowledge can be different for different persons.



Outline

Claim spaces

Analysis

Design

Evaluation

Providing more support

Conclusions

Analysis

An initial experiment to get some insight on the process of annotating a scholarly paper with its contributions and its connections to the literature.

A questionnaire, seven annotators, two documents, and a marker.

- q1: What is the problem tackled in this document ?
- q2: How does the work presented try to address this problem ?
- q3: What previous work does it build on ?
- q4: What previous work does it critique ?

Analysis

Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing

Katy Börner

Indiana University, School of Library and Information Science
10th Street & Jordan Avenue, Main Library 019, Bloomington, IN. 47405 USA
E-mail: katy@indiana.edu

ABSTRACT

The paper introduces an approach that organizes retrieval results semantically and displays them spatially for browsing. Latent Semantic Analysis as well as cluster techniques are applied for semantic data analysis. A modified Boltzman algorithm is used to layout documents in a two-dimensional space for interactive exploration. The approach was implemented to visualize retrieval results from two different databases: the *Science Citation Index Expanded* and the *Dido Image Bank*.

KEYWORDS: Digital Libraries, Browsing, LSA, Conceptual Clustering, Boltzman Algorithm, Information Visualization

INTRODUCTION

The wealth of digitally stored data available today increases the demand to provide effective tools to retrieve and manage relevant data. Keyword searches over digital libraries, repositories, or the Web easily retrieve lists of several hundreds of documents.

Information visualization - the process of analyzing and transforming data into an effective visual form - is believed to improve our interaction with large volumes of data.

First visual interfaces to digital libraries provided full-text searching and full-content retrieval capabilities and visualized documents according to authors, time, place, or citation relationships.

A considerable body of recent research applies powerful mathematical techniques such as *Factor Analysis*, *Multidimensional Scaling* or *Latent Semantic Analysis* to extract for example the underlying semantic structure of documents, the (evolving) specialty structure of a discipline, author co-citation patterns, changes in authors' influences in a particular field. In order to display the results of the data analysis spatially, computationally expensive techniques have to be applied to transform data analysis results to 2 or 3-dimensional coordinates. The computational expense of data analysis and visualization generation is very high. Therefore, precompiled, mostly static visualizations of fixed data sets are only displayed.

To our knowledge there exists no system that interactively visualizes retrieval results for browsing based on their underlying semantic structure.

DATA ANALYSIS

Latent Semantic Analysis (LSA) [4] has demonstrated improved performance over the traditional vector space techniques. It overcomes the problems of *synonymy* (variability in human word choice) and *polysemy* (same word has often different meanings) by automatically organizing documents into a semantic structure more appropriate for information retrieval. We apply LSA to extract the semantic structure of a particular database in a computationally expensive batch job.

At retrieval time, the result of a database query is hierarchically organized, based on the LSA output. Nearest neighbor-based, agglomerative, hierarchical, unsupervised conceptual clustering is applied to create a hierarchy of clusters grouping documents of similar semantic structure. Clustering starts with a set of singleton clusters, each containing a single document. The two clusters most similar are merged to form a new cluster that covers both. This process is repeated for each of the remaining clusters. At termination, a uniform, binary hierarchy of document clusters is produced. The partition showing the highest within-cluster similarity and lowest between-cluster similarity is selected for data visualization.

DATA VISUALIZATION

Rather than being a static visualization of data, the interface is self-organizing and highly interactive. Data is displayed in an initially random configuration, which sorts itself out into a more-or-less acceptable display via a modified Boltzman algorithm [1]. The algorithm works by computing attraction and repulsion forces among nodes based on the result of the data analysis. Nodes may represent articles or images which are attracted to other nodes to which they have a (reference or similarity) link and repelled by nodes to which there is no link. If the algorithm does not produce a visually acceptable layout, or if the user wishes to view the results differently, nodes can be grabbed and moved.

PROTOTYPE SYSTEMS

Two systems have been implemented in Java using the data organization and visualization methods described above.

SCI-E: The first system visualizes query results from the Science Citation Index Expanded (TM) as published by the Institute for Scientific Information®. The Citation Index provides access to current bibliographic information and cited references in more than 5,600 journals. Querying it via the Web of Science® Interface at http://webofscience.com/ results in an often huge number of matching documents organized in lists of ten that can be marked, saved, and downloaded for detailed study.

To demonstrate a visual browser to this kind of data base we will use DAIV188, a query result data set from SCI-EXPANDED that contains 188 articles matching the topic 'data AND analysis AND information AND visualization'.

The articles are represented in the usual Web of Science data output format (including author(s), article title and source, cited references, addresses, abstract, language, publisher information, ISSN, document type, keywords, times cited, etc.).

LSA was applied over keywords and abstracts of articles. As a result of conceptual clustering, the 167th partition was selected for visualization. It contains 20 clusters grouping 1 - 53 articles. Figure 1 shows the Java interface. Each book article is represented by a rectangle and each journal article by an oval. Articles are labeled by their first author. Lines between nodes visually represent co-citation links.

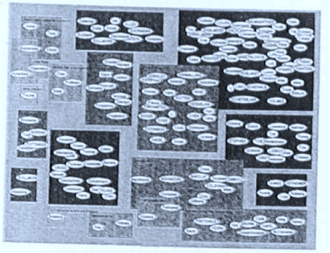


Figure 1: Java Interface to DAIV188

The 2-dimensional layout of articles corresponds to the data mining result as well as to the forces applied by the Boltzman algorithm to generate an acceptable layout. The higher the similarity of articles within a cluster the lighter its color. Each cluster is labeled by the keyword used most often.

DIDO: Another instantiation of the system enables users to browse search results from the Dido Image Bank, http://www.dlib.indiana.edu/collections/dido/ provided by the Department of the History of Art, Indiana University. Dido stores about 9,500 digitized images from the Fine Arts Slide Library collection of over 320,000 images. Each image in Dido is stored together with its thumbnail representation as well as a textual description. LSA was applied over the textual descriptions exclusively. For demonstration purposes the set of images matching the keyword descriptor 'MONET' were retrieved and displayed for browsing. It contains 21 documents inclusive two portraits of Claude Monet drawn by Edouard Manet (see Figure 2).

Figure 2: The MONET Clu

Thumbnail representations of images hav
the Dido Database showing some of Mo
such as haystacks, cathedrals, and water lil

CONCLUSIONS

Initial tests show that the presented appr
access to textual materials, such as arti
documents for which textual descriptions a
images. Detailed user studies are in prepara
First results on using an immersive 3-
environment for the interactive exploratio
are presented in [3].
An extended version of this paper as well
versions of Figures 1 and 2 a
http://ella.slis.indiana.edu/~katy/DL00.

ACKNOWLEDGMENTS

Robert Goldstone, Mark Steyvers, Helen
Fry have been valuable discussion partner
[2] by M. Berry was used for computing
decomposition. The research is suppor
Performance Network Applications grant o
are Andrew Dillon and Margaret Dolinsky.

REFERENCES

1. Alexander, Garcia, and Alder. Simulati
Consistent Boltzman Equation for Hard
Extension to Dense Gases, *Lecture Not*
Springer Verlag, 1995.

2. Berry, M. et al. SVDPACKC (Version
University of Tennessee Tech. Report C
(Revised October 1996).

3. Börner, K. Visible Threads: A smart VR
digital libraries. *Electronic Imaging 200*
Exploration and Analysis.

4. Landauer, T. K., Foltz, P. W., & Laham
to Latent Semantic Analysis. Discourse
259-284, 1998.

Analysis

#	Document component	a1	a2	a3	a4	a5	a6	a7
	Extracting. . . [title]							q1
	Abstract							
1	The paper introduces. . .	q2	q2	q1	q2	q2 q3	q1	
2	Latent Semantic. . .	q2	q3	q2		q2 q3		q1
3	A modified Boltzman. . .	q2	q3	q3		q2 q3		
4	The approach was. . .	q3						
	Keywords							
	Introduction							
5	The wealth of. . .	q1		q1	q1	q1		
6	Keyword searches over. . .		q1		q1			
. . .	. . .							

Analysis

Observations:

- Different answers to a question
- Different components of the article used to answer the same question
- A component can be used to answer different questions

Components: paragraphs, sentences, clauses, verbs, metadiscourse. . .

"Interpreting" as "positioning oneself with respect to the author's stance".

A **suggestion** approach: identify and recommend to the annotator a set of elements from the text and/or the repository of claims and propose them for consideration.

Analysis

The sources of information we can consider include:

- Content-based extracted elements
- Scaffolding
- Peers' annotations
- Connected and related documents

We are aiming at integrating reading and annotating in a single process:

- Contextualising reading by displaying the suggestions in situations (where applicable)

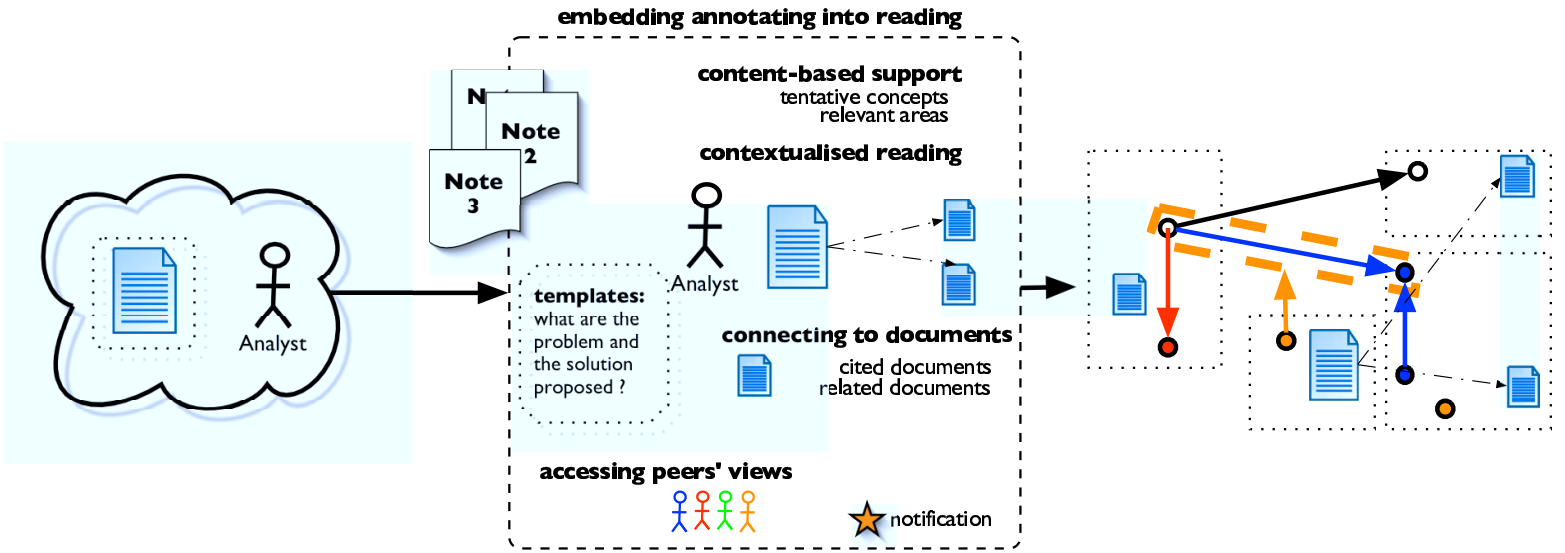
Analysis

Content-based extracted elements can be used to get some insight in the author's point of view:

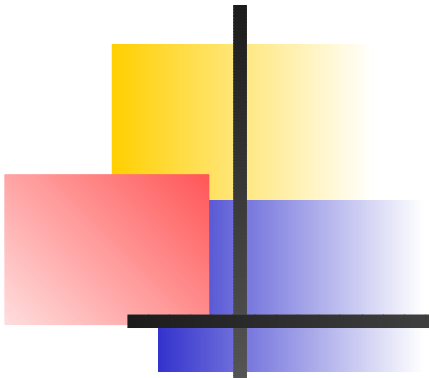
- Tentative concepts: keywords, most frequent noun groups
- Tentative 'author-made claims': instances of the relations defined in the ontology. Selected synonyms.
- Areas to could focus on: scholarly article components (sections, paragraphs, figures, keywords), 'important' sentences, rhetorically-coherent zones [Teufel & Moens, 2002]

Analysis

Supporting the transition from a scholarly document to a network of claims:



Document annotation is integrated into the reading process and supported by the manipulation of the source document.



Outline

Claim spaces

Analysis

Design

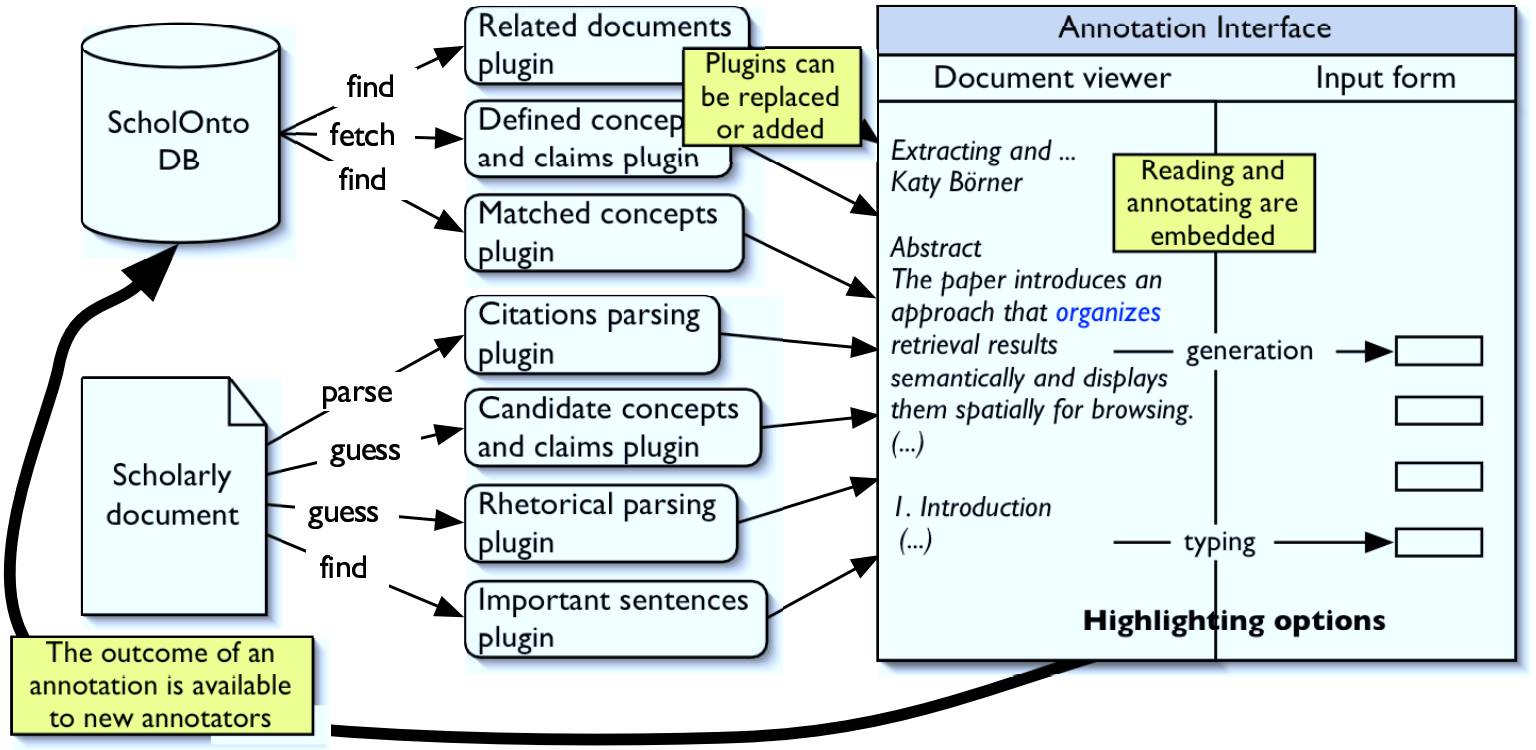
Evaluation

Providing more support

Conclusions

Design

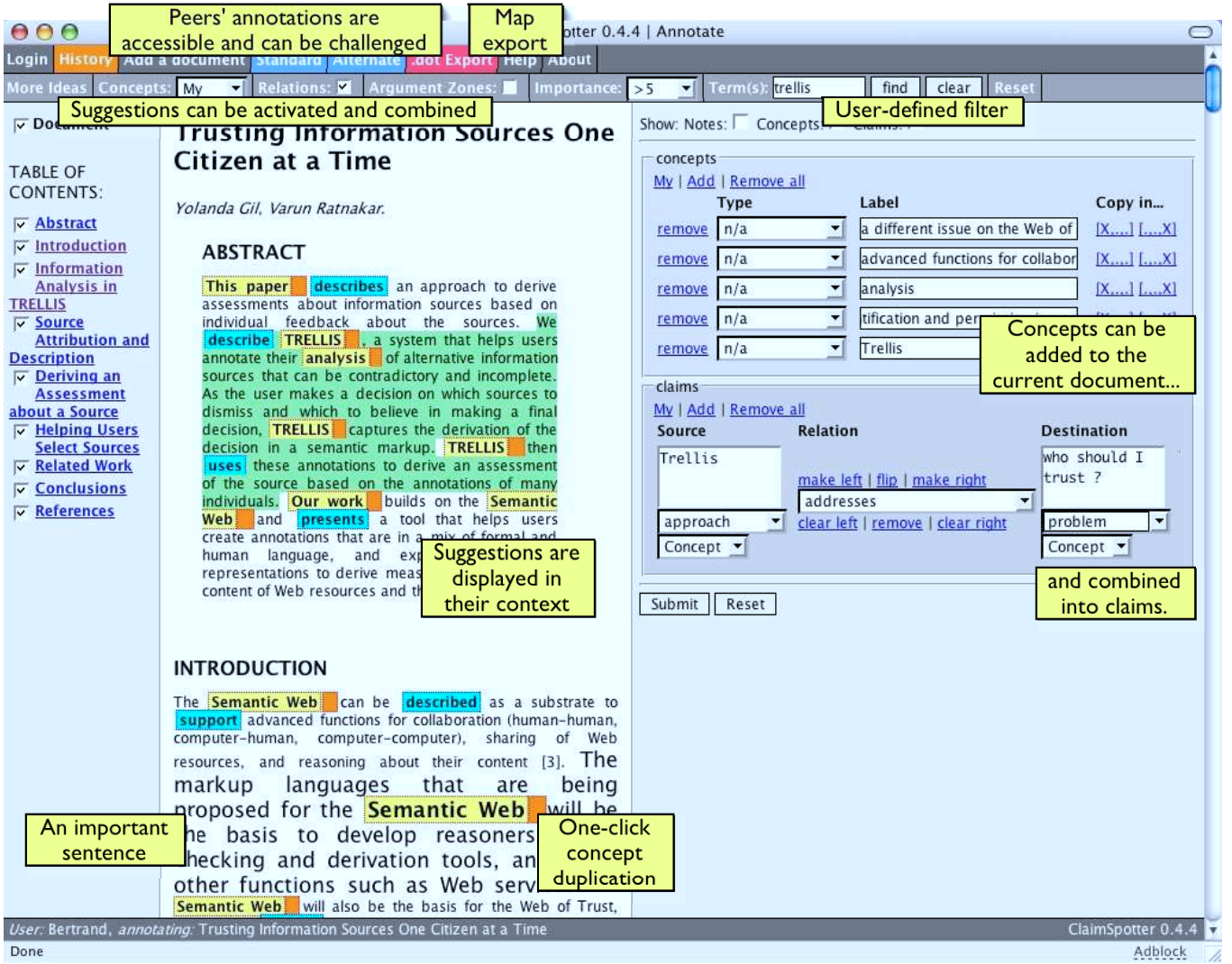
Architecture:



Plugins communicate their results with XML.

This should allow their replacement with more robust iterations.

Design


Peers' annotations are accessible and can be challenged
Map export
Suggestions can be activated and combined
User-defined filter
Concepts can be added to the current document...
Suggestions are displayed in their context
and combined into claims.
An important sentence
One-click concept duplication
Trusting Information Sources One Citizen at a Time
Yolanda Gil, Varun Ratnakar.
ABSTRACT
INTRODUCTION
TABLE OF CONTENTS:
Abstract
Introduction
Information Analysis in TRELLIS
Source Attribution and Description
Deriving an Assessment about a Source
Helping Users Select Sources
Related Work
Conclusions
References
concepts
claims
Submit
Reset

Design

Jim Blythe on ‘Task Learning by Instruction in Tailor’:

My concepts: Tailor, a system that allows users to modify task information through instruction, an evaluation to assess whether the system makes users’ life easier, User training. . .

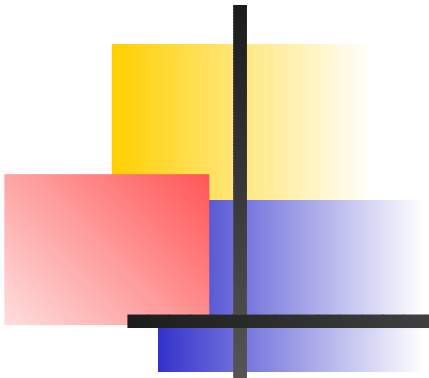
My claims: [Tailor, is about, Making intelligent systems more widespread], [Tailor, addresses, lack of flexibility in systems descriptions]. . .

(movie)

Design

After this annotation is finished:

- This document's concepts and claims become available (for consultation or reuse) to further annotators. Annotators can look at previously added claims about the document and take position with them.
- Discover related documents: documents sharing a concept, or being connected in a claim
- The set of concepts being in an 'addresses' claim ending with 'lack of flexibility in systems descriptions' is updated.
- Claim authors get notified if one of their claim is attacked



Outline

Claim spaces

Analysis

Design

Evaluation

Providing more support

Conclusions

Evaluation

A formative evaluation based on observation studies.

Experimental protocol

- 13 participants (9 beginners, 4 experts)
- **Task:** annotate a paper they were familiar with its contributions and the connections that either the paper's author was making or that the annotators wanted to make.
- Audio and video recordings. Cooperative evaluation with an expert to assist them in the process. These were transcripted and analysed qualitatively.
- 1 hour for each session

Evaluation

How do we know we are on the right track ?

Participants managed to get something done within the allocated time.

What worked ? What didn't ?

The ability to modify a document and reduce it to as little or as much as wanted has been appreciated. Possibilities to access the history and reuse previous annotations in a click were also welcomed.

The amount of information proposed was sometimes overwhelming.

The problem is 'do you make your own claims, do you follow the system, do you go back to the history to see what the other people have said'

Are our sources of support useful ? And used ?

As many behaviors as participants. On average, they made a decent use of the suggesting filters. Some extremes cases.

Evaluation

Sub-task #1: identifying concepts

Annotation starts with creating a few concepts.

Is the environment influencing the formulation of these concepts ?

Potentially, yes (but similar concepts could maybe have been formulated without it).

The highlighting of these concepts in the text can shape.

Reusing an existing concept even if it is not exactly what one wanted.

It is less expensive.

Quoted: "I want to ‘add some color in there".

Evaluation

Sub-task #2: articulating concepts into claims

Starting from the relation vs. 'starting from the concepts to connect'

'I want to combine concepts with an 'addresses' relation' vs. 'This concept and this one are connected'.

A difference between experts and beginners. Less flagrant over time.

Reformulating a concept or a claim to make it fit the formalism

Formalizing is translating, id est losing a part of the original meaning

Switching left and right parts ('I am just throwing concepts in') can help

Evaluation

Sub-task #2: articulating concepts into claims (cont'd)

Reformulations

'I want to say "A limitation of Magpie is that it is not able to use existing semantic annotations". So I would like to add a claim. I put (the concept) 'Magpie' on the left hand side, and then see. . . Actually, you cannot say. . . You have to say in a different way. . . You have to create a concept 'Inability to use existing semantic annotations' (. . .) and another one 'Problem with Magpie' and connect them with 'is an example of' -> 'Inability to use existing semantic annotations', 'is an example of', 'problems with Magpie'.

Evaluation

Open questions:

- Is there too much information available ?
- What to say ? Where to stop ? (granularity, boundaries)
- Guiding towards existing relations
  - Annotators have to ‘subscribe’ to the underlying formalism
  - There is nothing one can do if they want to say something that is not captured by it
  - However, we can try to make them say more ’ScholOnto-compatible’ things and less ’general’ things

Outline

Claim spaces

Analysis

Design

Evaluation

Providing more support

Conclusions

Providing more support

Scaffolding:

q1	What is the problem identified in this document ?
q2	How is this problem related to other problems ?
q3	What are the proposed approach and solution ?
q4	What are the claims connecting problem and solution ?
q5	How is this solution related to other solution(s) proposed to address this problem ?

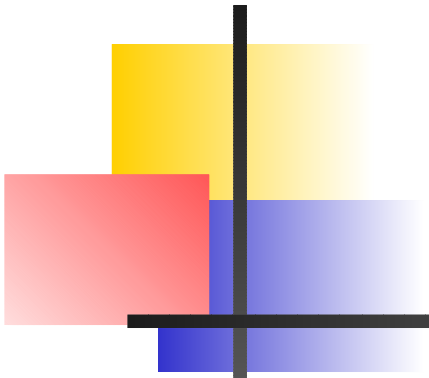
Support to answer q1 (problem): sentences classified as AIM by a rhetorical classifier, concepts defined over this document which have been typed as ‘problem’, and the destination ends of claims using an *addresses* relation.

Providing more support

Mini-evaluation with two participants (one expert and one one beginner).

Initial remarks:

- as a "*walkthrough/overview (especially if I am not totally familiar with the document)*", or to "*make you think about a paper and identify its structure*.
- "*I wouldn't want to be flooded with everyone else answers.*" (filtering options)
- there will be times where she would "*spend time making my own claims.*" Conversely, "*there would (also) be times where extensive reuse of claims is the best approach.*"



Outline

Claim spaces

Analysis

Design

Evaluation

Providing more support

Conclusions

Conclusions

Supporting the annotation of scholarly documents is a tough design problem.

- Understanding the issues involved in the annotation of a scholarly document with claims
- A supportive UI based on the idea of suggestions to reduce the information overload
- An empirical study which identified positive and negative aspects of the approach.

Lessons learnt:

- Users will have different needs based on the amount of time they want to put, or the relative importance of the paper.
- More ‘intelligent support’ is needed to provide guidance to the annotators.
- The way the information is presented influences the modeling process (concepts)

Conclusions

Future work:

- More filters.
- Support to answer the questions
- Integration with a sketching environment to model one's interpretation

Appendix

`http://kmi.open.ac.uk/projects/scholonto`

Relations discourse ontology

Category	Instance
general	is about, uses/applies/is enabled by, improves on, impairs
problem	addresses, solves
supports	proves, refutes, is evidence for, is evidence against, agrees with, disagrees with, is consistent with, is inconsistent with
taxonomic	part of, example of, subclass of
similarity	is identical to, is similar to, is different to, is the opposite of, shares issue with, has nothing to do with, is analogous to
causal	predicts, envisages, causes, is capable of causing, is prerequisite of, is unlikely to affect, prevents

Statistics

Most used relations:

uses/applies/is enabled by: 18.2%

is about: 13.8%

example of: 8.8%

addresses: 7.5%

is evidence for: 7.50%

Beginners used more ’is about’ than experts.

More talkative annotators (submitting 10+ claims) used ’is about’ in 15.2%; less talkative ones in 4.5%