# ClaiMaker: Weaving a Semantic Web of Research Papers

**Gangmin Li, Victoria Uren, Enrico Motta,**
**Simon Buckingham Shum, John Domingue**

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK
{g.li, v.s.uren, e.motta,
s.buckingham.shum, j.b.domingue}@open.ac.uk
http://kmi.open.ac.uk/projects/scholonto/

**Abstract.** The usability of research papers on the Web would be enhanced by a system that explicitly modelled the rhetorical relations between claims in related papers. We describe ClaiMaker, a system for modelling readers' interpretations of the core content of papers. ClaiMaker provides tools to build a Semantic Web representation of the claims in research papers using an ontology of relations. We demonstrate how the system can be used to make inter-document queries.

## 1 Introducing ScholOnto

The Web has facilitated access to many scholarly documents by making available copies of papers, technical reports etc. in digital libraries and on individuals' home pages. Reasonable keyword access is provided by Web search engines. Access via citations is available using tools such as Research Index (Citeseer) [1], and research to extend this approach to eprint servers is ongoing [2]. However, there are few tools to track debate and analyse ideas in a domain. The Semantic Web [3] approach of augmenting Web documents with machine understandable information offers a potential means of addressing this need.

The Scholarly Ontologies (ScholOnto) project [4, 5] takes this approach. We are developing an ontology-based *Claims Server* to augment existing papers, by modelling authors' and readers' interpretations of them. This produces a *claim space* above raw digital libraries; effectively, a semantic web of inter-linked concepts. The system enables researchers to make claims concerning their view of a document's contributions and its relationship to the literature. These claims may support or contest existing claims; in contrast to most Semantic Web applications ScholOnto does not require consensus.

The semantic structure of the claim space provides a basis for making queries based on the interpretation of research papers, rather than just keywords or citations. In this paper, we consider one example of an apparently simple question, which requires interpretation of multiple documents in a more specific way than is possible from plain citations: *"Are there any arguments against the intellectual framework on which this paper builds?"*. We will show how building a semantic network of claims over a distributed document collection can start to answer such questions.

## 2 Ontology of Rhetorical Relations

We take the position that, although *what* authors are discussing in a domain will, by the nature of research, be in flux, *how* the discourse is conducted will be stable. Consequently, the conceptual glue of ScholOnto, the links between ideas, is reified using an ontology of rhetorical relations [6]. A claim triple is the assertion that a

particular relationship holds between two ideas. The relations in the ontology act as attributes in triples, in which object and value are each one of concept, set or data. Concepts are stored as short pieces of free text, and sets are collections of related concepts gathered under a free text name. A typical data object is a set of metadata giving the reference of a document in a digital library.

Claims were modelled in a range of research domains, including computer supported collaborative work, text categorization, and literary criticism. Relations common to several domains were identified. We found we could classify these into groups with similar rhetorical implications: Supports/Challenges, Problem Related, Taxonomic, Causality, Similarity, and General. Each relation belongs to one group. We also found that some relations occurred in pairs of opposites, e.g. *proves* and *refutes*, where one has positive and the other negative implications. We call this property "polarity". For example, *refutes* has negative polarity; it implies *dis*proof. Referring to our question, *refutes* would be an *"argument against"*.

```
:SchProperty rdfs:subClassOf :Property .
:StructuringProperty rdfs:subClassOf :SchProperty .
:RhetProperty rdfs:subClassOf :SchProperty .
:SupportsChallenges rdfs:subClassOf :RhetProperty .

:PolarityType rdf:subClassOf :Resource .
:negativePolarity rdf:type :PolarityType .
:positivePolarity rdf:type :PolarityType .

:polarity rdf:type :StructuringProperty .
:polarity rdfs:domain :SchProperty .
:polarity rdfs:range :PolarityType .

:refutes rdf:type :SupportsChallenges .
:refutes :polarity :negativePolarity .
:proves rdf:type :SupportsChallenges .
:proves :polarity :positivePolarity .
```

**Fig. 1.** Parts of an RDFS specification for the ScholOnto ontology (in Notation3 for clarity http://www.w3.org/DesignIssues/Notation3)

By defining relations in terms of type and polarity we can reason with them at a higher level of granularity than individual relations; it is not just the claims made using the *refutes* relation that represent *"arguments against"* something, but any claims made using links that have negative polarity. Furthermore, the same ontology of relations can be employed by research communities which speak different "dialects", or even different languages, simply by changing the labels of the relations, without changing the underlying functionality of ScholOnto.

To illustrate claim triples, we will take a paper entitled "Evaluation of decision forests on text categorization" [7]. The claims of this paper include the following:

[Decision Forest Classifier] (uses/applies/is enabled by) [Decision tree learning]
This uses one of the General relations *uses*/*applies*/*is enabled by* to assert that the *Decision Forest classifier* studied in the paper uses a well known method, *Decision tree learning*. The latter concept was introduced in a different document, so this link has a contextual role: it locates the paper near similar claims.

[Decision Forest classifier improves on C4.5 and kNN] (is inconsistent with) [SVM and kNN outperform other classifiers]
This claim uses the negative, Supports/Challenges relation *is inconsistent with* to link one of the experimental results of this paper to a result in a third paper. In addition to its contextual role, locating the claim near other comparisons of classifiers, this claim has a rhetorical role: it contrasts pieces of evidence that make contradictory assertions.

## 3 The ClaiMaker System

ClaiMaker is implemented as a client/server system (Fig. 2). The Claim Server interprets users' requests, and accesses the database and/or file server to retrieve the results. It may invoke the inference engine, based on the relation ontology, if it is necessary.

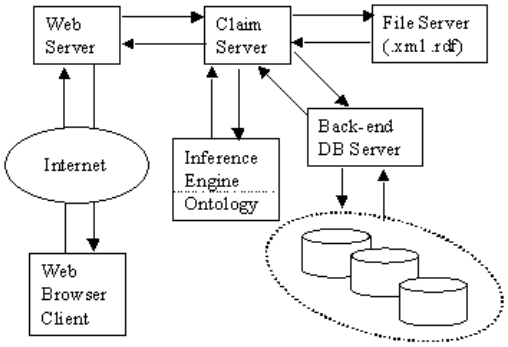


**Fig. 2.** Architecture of the ClaiMaker Claim Server

ClaimMaker has a form-based interface to help ourselves and early uptake users build a claim space, which describes a collection of electronic documents. The operations it performs include: adding or importing metadata for new documents; creating new concepts, sets and links associated with a document; and browsing and querying the database about objects on the server to discover interesting facts, and potential trends. The interface leads the user through ScholOnto tasks stepwise. For example, Figure 3 shows a user selecting concepts to include in a set about reminding.

| ID | IPOwner | Creator | Article | Concept name | Select |
|---|---|---|---|---|---|
| 644 | Victoria | Victoria | 2464 | Reminding and Memory | ☐ |
| 645 | Victoria | Victoria | 8707 | Reminding is a crucial aspect of human memory | ☑ |
| 647 | Victoria | Victoria | 8707 | Events can remind you of events in the same domain | ☐ |
| 648 | Victoria | Victoria | 8707 | Events can remind you of events in different domains | ☐ |
| 652 | Victoria | Victoria | 8707 | Understanding means being reminded of the closest previously experienced phenomenon. | ☐ |
| 663 | Victoria | Victoria | 8707 | Reminding can show how memory is organized | ☑ |
| 664 | Victoria | Victoria | 8707 | Reminding tells us about learning and generalization | ☑ |

Selection is done

**Fig. 3.** Selecting concepts to construct a set

In Figure 4 the user is making a claim using this set, which they have named "Importance of Reminding" and the relation *is consistent with*. The next step will be to click the button *Search concept/set* which will take them into a screen where they can make keyword searches of other users' concepts and sets, and select one to link to.

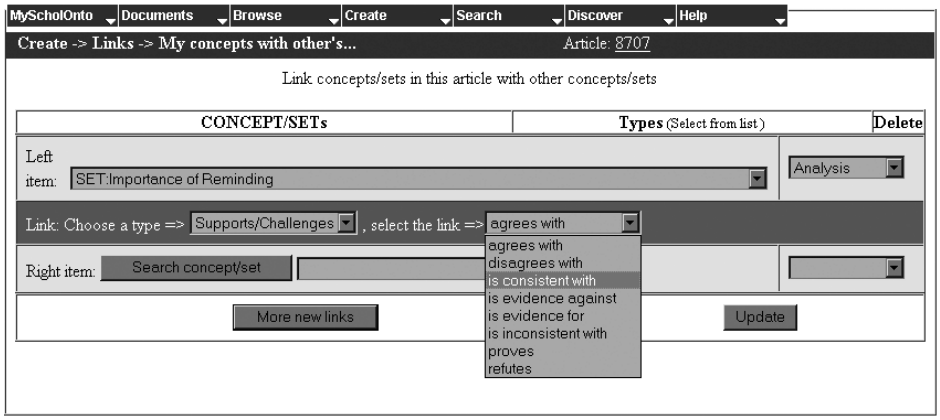


**Fig. 4.** Creating a claim using the ClaiMaker system

## 4 Providing Semantic Discovery Services

We will now return to our example query to demonstrate how expressing the claims made by documents using the ontology of relations gives added value over retrieval of documents based on keywords. The question as asked, *"Are there any arguments against the intellectual framework on which this paper builds?"*, has three components. It is looking for *"arguments against"*, defined as negative relations of any type. It refers specifically to a *"paper"*, and it is easy to find the set of concepts belonging to a document. It also refers to the *"intellectual framework"*. This is an ambiguous requirement that must be constrained if it is to be modelled. For the experimental function described here, we used a pragmatic definition: *the intellectual framework of a set of concepts is the extended set of concepts that are linked to/from the concepts in the original set by a positive relation*. Clearly, this is a gross simplification of the notion of *"intellectual framework"*, but it makes the problem tractable.

For a given paper the discovery function does the following:
1. Finds the concepts associated with that paper
2. Extend the set of concepts by adding linked concepts from other papers
3. Returns any arguments against the concepts in the extended set

Typical results are presented below (Fig. 5). Note the two numbers to the right of the claim that *disagrees with* one of the related issues in the query. The first, 8621, is a hyperlink to the metadata of the paper that provides the backing for the claim, which includes a URL to the paper itself. The second, 2, is a link to the personal details of the reader who made the claim; this allows the user, or, potentially, a discovery agent working on behalf of the user, to make a judgement about the credentials of a claim; can it be trusted?

| The key issues you are concerned with: | |
|---|---|
| 445 | Decision Forest classifier |
| 446 | Decision Forest classifier improves on C4.5 and kNN |

| The related issues you may be concerned with: | |
|---|---|
| 515 | Instance based learning |
| 511 | Decision tree learning |
| 277 | decision trees and naive Bayes perform well for text categorization |

| The following claims disagree ... | | | |
|---|---|---|---|
| 1 | [Optimised rules outperform Naive Bayes and decision trees] «disagrees with» [decision trees and naive Bayes perform well for text categorization] | 8621 | 2 |

**Fig. 5.** Arguments that contrast with the concepts in the paper by Chen & Ho [7]

Term based information retrieval handles documents as isolated entities defined by the words in them. Citations in a document are noncommittal about authors' intentions in referring to other work; we cannot even tell if a paper is referenced because the authors support its position or because they are diametrically opposed to it. This simple example of a search for arguments against a position demonstrates how the ontology of relations can make the connections between ideas in different documents explicit, allowing better kinds of query.

## 5 Summary & Future Work

The ontology we have implemented in ScholOnto permits us to represent researchers' claims about their work as a claim space over Web documents. This opens up opportunities for answering more interesting questions about scholarly discourse.

We are now developing more discovery services. These will be of two types. We will start by developing specific functions of the sort discussed here. These will tackle common tasks, like finding the arguments against a position, or assessing the impact of an idea. Novice users will be able to use these to learn about the sorts of query possible in ScholOnto. In addition, we plan to develop a structural query system, exploiting the inference engine. This system will be aimed at expert users.

Data visualisation will become increasingly important. We need visualisations for browsing that illustrate the claim space at different levels of granularity. A visual input system is required also. When making a list of claims it is easy to lose track of the shape of the argument that is being made, and how it relates to other parts of the network. Users need to be able to see the connections between their claims as they create them. We are also investigating ways to extract claims from papers semi-automatically, and to suggest semantic links, as a way of easing the claim acquisition bottleneck.

The Claims Server implementation described here provides a controllable, centralised environment in which we can test our ideas. However, an agent approach [8] offers some exciting alternatives. One is a distributed ScholOnto in which authors' interpretations of their own papers are published alongside the originals. These could be perused by discovery agents. Another is a more personalised model in which a user's agent might crawl the Web, harvesting interesting claims as they are published, and depositing them in a private knowledge base. They could then be annotated and extended, without the social constraints imposed by making claims about other researchers' work in public. Such private spaces could be shared by the members of a research group as a discussion forum.

## References


1. Bollacker,K.D., Lawrence, S., Giles, C.L.: CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. Proc. 2nd Int. Conf. on Autonomous Agents, Minneapolis, MN (1998) 116-123
2. Hitchcock, S., Carr, L. Jiao, Z., Bergmark, D., Hall, W., Lagoze, C., Harnad, S.: Developing services for open eprint archives: globalisation, integration and the impact of links. In: Proc. 5th ACM Conf. on Digital Libraries, San Antonio, TX. (2000) 143-151
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific American, May (2001) 34-43
4. Motta, E., Buckingham Shum, S., Domingue, J.: Ontology-Driven Document Enrichment: Principles, Tools and Applications. Int. J. Human-Computer Studies., 52, (2000) 1071-1109
5. Buckingham Shum, S., Motta, E., Domingue, J.: ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse. Int. J. Digit. Libr., 3, (2000) 237-248
6. Buckingham Shum, S., Uren, V., Li, G., Domingue, J., Motta, E., Mancini, C.: Designing Representational Coherence into an Infrastructure for Collective Sensemaking. In: 2nd Int. Workshop on Infrastructures for Distributed Collective Practices, San Diego CA (2002)
7. Chen, H. & Ho, T.K.: Evaluation of decision forests on text categorization. In: Proc. 7th SPIE Conference on Document Recognition and Retrieval (2000) 191-199
8. Hendler, J.: Agents and the Semantic Web. IEEE Intelligent Systems, 16(2) (2001) 30-37