Whetting the Appetite of Scientists: Producing Summaries Tailored to the Citation Context

Stephen Wan, Cécile Paris
ICT Centre
CSIRO
Sydney, Australia
{Stephen.Wan|Cecile.Paris}@csiro.au

Robert Dale
Centre for Language Technology
Faculty of Science
Macquarie University, Australia
rdale@science.mq.edu.au

ABSTRACT

The amount of scientific material available electronically is forever increasing. This makes reading the published literature, whether to stay up-to-date on a topic or to get up to speed on a new topic, a difficult task. Yet, this is an activity in which all researchers must be engaged on a regular basis. Based on a user requirements analysis, we developed a new research tool, called the *Citation-Sensitive In-Browser Summariser* (CSIBS), which supports researchers in this browsing task. CSIBS enables readers to obtain information about a citation at the point at which they encounter it. This information is aimed at enabling the reader to determine whether or not to invest the time in exploring the cited article further, thus alleviating information overload. CSIBS builds a summary of the cited document, bringing together metadata about the document and a citation-sensitive preview that exploits the citation context to retrieve the sentences from the cited document that are relevant at this point. This paper briefly presents our user requirements analysis, then describes the system and, finally, discusses the observations from an initial pilot study. We found that CSIBS facilitates the relevancy judgment task, by increasing the users' self-reported confidence in making such judgements.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific databases; H.5.2 [**User Interfaces**]: Natural language, User-centered design; H.5.4 [**Hypertext/ Hypermedia**]: Navigation, User issues

General Terms

Design, Human Factors

Keywords

Information needs; Information browsing; Scientific Literature; Biomedical Researchers; User Modeling and Interactive IR; Summarization

JCDL'09, June 15–19, 2009, Austin, Texas, USA.
Copyright 2009 ACM 978-1-60558-322-8/09/06 ...$5.00.

1. INTRODUCTION

Researchers regularly browse through repositories of on-line academic literature to update their existing knowledge or to quickly familiarise themselves with a new topic. In a survey we conducted of biomedical researchers, for example, we found that two-thirds of participants browsed the academic literature at least once a week. Like many knowledge workers, they are increasingly time poor. That, coupled with a near exponential growth of available publications and scientific material, makes the task of browsing through and staying up-to-date with academic literature a particularly onerous one. With the growing body of on-line material, there is both an opportunity and a need to develop new tools to support researchers in this browsing task and to help them make decisions about the relevance of articles.

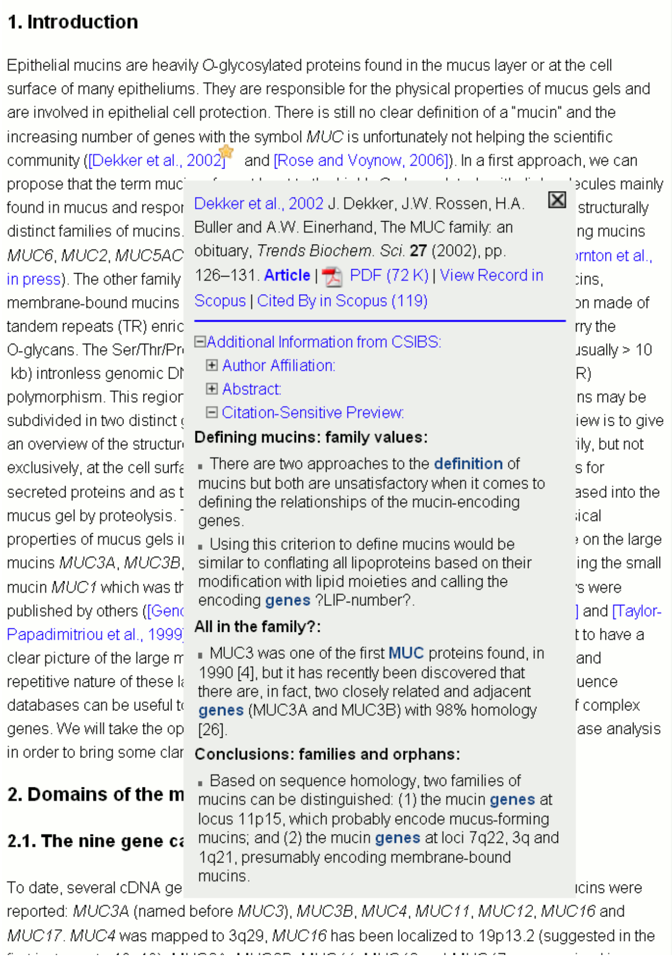
There are typically two means of finding documents that address the researcher's information needs: searching and browsing. By *searching*, we might find relevant material through queries posted to search engines. We also regularly expect to find additional important information by *browsing* through documents and following potentially relevant citations.

As one reads a document (the *citing* document), a large number of citations may be encountered. Each citation points to a *cited* document and is embedded within a *citation context*. While some tools (for example, *ScienceDirect*[1]) enable the reader to obtain, at the point of citation, the full details of the bibliographic reference, this is often insufficient information to determine if the referred-to work is worth reading in full. This is the problem we address here: we want to support researchers in determining whether a cited document is relevant to their needs and whether they should go to the trouble and cost of obtaining it. We present the *Citation-Sensitive In-Browser Summariser* (CSIBS), a tool that supports readers in deciding which cited documents they should read. CSIBS was developed based on a user requirements analysis, which revealed two prominent tasks: appraising the cited document and determining if the citation was justified. Ultimately, in performing these two actions, the reader is trying to decide whether or not to read the cited document in full, and it is this overall task that CSIBS strives to facilitate.

CSIBS helps readers browse and navigate through a dense network of cited documents. It does this by presenting to readers, at the point of citation, a summary of the cited document. The summary contains both important metadata about the cited document (e.g., the author's affiliations

[1] See `www.sciencedirect.com`.

or the impact factor of the host journal) and key sentences extracted from the cited document, *based on the citation context* (or the reader's reading context).



1. Introduction

Epithelial mucins are heavily O-glycosylated proteins found in the mucus layer or at the cell
surface of many epitheliums. They are responsible for the physical properties of mucus gels and
are involved in epithelial cell protection. There is still no clear definition of a "mucin" and the
increasing number of genes with the symbol *MUC* is unfortunately not helping the scientific
community ([Dekker et al., 2002] and [Rose and Voynow, 2006]). In a first approach, we can
propose that the term muc ecules mainly
found in mucus and respor structurally
distinct families of mucins. ng mucins
MUC6, *MUC2*, *MUC5AC* rnton et al.,
in press). The other family cins,
membrane-bound mucins on made of
tandem repeats (TR) enric rry the
O-glycans. The Ser/Thr/Pr usually > 10
kb) intronless genomic D R)
polymorphism. This region ns may be
subdivided in two distinct iew is to give
an overview of the structur ily, but not
exclusively, at the cell surfa s for
secreted proteins and as t ased into the
mucus gel by proteolysis. ical
properties of mucus gels i on the large
mucins *MUC3A*, *MUC3B*, ing the small
mucin *MUC1* which was th s were
published by others ([Gen] and [Taylor-
Papadimitriou et al., 1999 t to have a
clear picture of the large m and
repetitive nature of these l uence
databases can be useful t f complex
genes. We will take the op ase analysis
in order to bring some clar

2. Domains of the m

2.1. The nine gene c

To date, several cDNA ge icins were
reported: *MUC3A* (named before *MUC3*), *MUC3B*, *MUC4*, *MUC11*, *MUC12*, *MUC16* and
MUC17. *MUC4* was mapped to 3q29, *MUC16* has been localized to 19p13.2 (suggested in the

Dekker et al., 2002 J. Dekker, J.W. Rossen, H.A. Buller and A.W. Einerhand, The MUC family: an obituary, *Trends Biochem. Sci.* **27** (2002), pp. 126–131. **Article** | PDF (72 K) | View Record in Scopus | Cited By in Scopus (119) ☒

⊟Additional Information from CSIBS:

⊞ Author Affiliation:

⊞ Abstract:

⊟ Citation-Sensitive Preview:

Defining mucins: family values:

- There are two approaches to the **definition** of mucins but both are unsatisfactory when it comes to defining the relationships of the mucin-encoding genes.
- Using this criterion to define mucins would be similar to conflating all lipoproteins based on their modification with lipid moieties and calling the encoding **genes** ?LIP-number?.

All in the family?:

- MUC3 was one of the first **MUC** proteins found, in 1990 [4], but it has recently been discovered that there are, in fact, two closely related and adjacent **genes** (MUC3A and MUC3B) with 98% homology [26].

Conclusions: families and orphans:

- Based on sequence homology, two families of mucins can be distinguished: (1) the mucin **genes** at locus 11p15, which probably encode mucus-forming mucins; and (2) the mucin **genes** at loci 7q22, 3q and 1q21, presumably encoding membrane-bound mucins.

Figure 1: A sample pop-up with an automatically generated summary. This pop-up appears when the user mouses over the citation. Words that match the citation context are coloured and emboldened. The figure also illustrates the presentation of relevant section titles, used to group extracted sentences. This grouping provides better organisation of the citation-sensitive preview.

The key observations here are that, while the meta-data about the document might enable the reader to appraise the cited document, the citation-sensitive preview (the portion of the summary sensitive to the reading context) provides information that can support the reader in determining whether the citation is justified in its context. Our hypothesis is that the content of this citation context can allow CSIBS to determine what information in the cited document will be of most value, given what has just been read by the reader in the citing document. This portion of the summary is built using automatic text summarisation methods that exploit the links and anchor texts in hypertext documents [14].

The summary is shown as a pop-up text box within the same browser in which the citing document is being viewed (for example, Adobe Acrobat Reader or a web browser). The integration of this support into the browser, by overlaying summaries on top of the citing document, enables readers to maintain their focus of attention by providing relevant snippets tailored to their information needs. In doing so, CSIBS aims at helping to avoid following irrelevant citations. An added advantage is that the original reading task is not disrupted.

Viewing the summary within a ScienceDirect webpage is triggered by moving the mouse over the hyperlink to the cited document. Within the PDF version, the summary is similarly activated either by moving the mouse over a citation or by double-clicking on the citation itself. The former simply provides the summary, while the latter allows the user to copy and paste the generated summary. To provide these interaction modes, CSIBS exploits functionality provided in the implementation of Javascript pop-ups within the Adobe Acrobat Reader. Examples of this pop-up mechanism are shown in Figures 1 (a web version) and 2 (a PDF version). The work described in this paper is currently demonstrated on publication data served by databases maintained by Elsevier.[2]

The present paper is structured as follows. We first present related work in Section 2. Section 3 briefly describes the user study we carried out to determine the primary information needs of readers as they encountered citations; this study was carried out in the biomedical domain. Section 4 describes the design and implementation of the CSIBS system, which was based on the results of our user study. Section 5 discusses our preliminary evaluation study. Finally, Section 6 presents our current on-going and future research interests, while Section 7 concludes the paper.

2. RELATED WORK

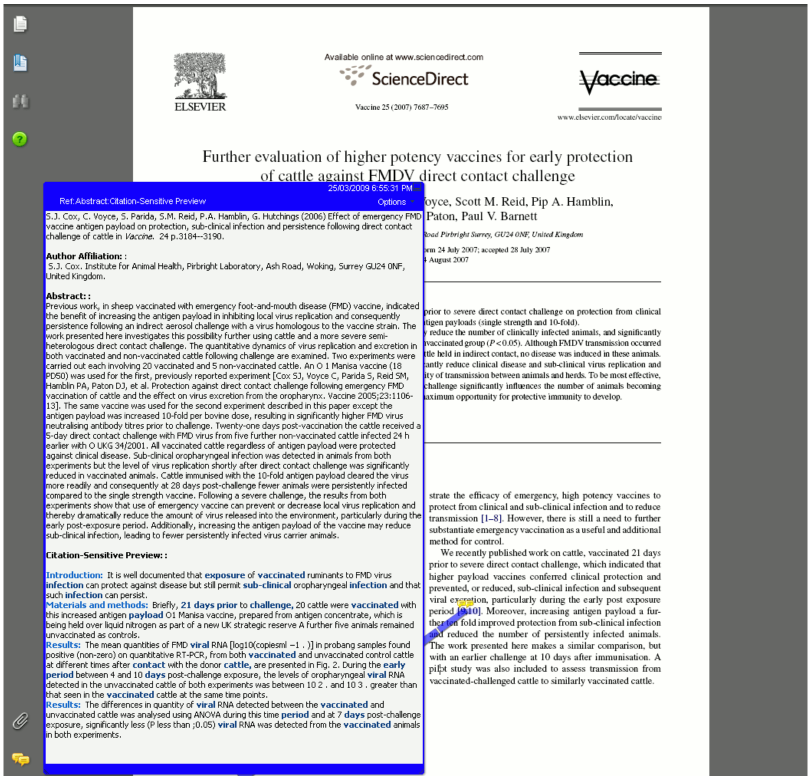
The tasks of users in academia, specifically in the humanities [1] and in mathematics [15], have been studied previously, revealing a number of teaching and research activities that might be supported by technology. In this work, we are specifically interested in supporting the task of making a relevance judgement about a cited document. In particular, CSIBS is aimed at supporting users in the science disciplines.

Li et al. [3] analysed the usage of scientific text, specifically focusing on the types of queries made when searching over large-scale scientific literature repositories. Our analysis differs in that we seek to understand the user's scenario in terms of information needs and the underlying tasks being performed, particularly in the context of browsing, as opposed to searching, through literature.

To facilitate these tasks, we investigate technologies that are able to locate and present useful information from within academic documents. Other research has addressed similar goals: the extraction of tabular information from within scientific documents has been explored [5]; and key phrases from scientific publications are identified in [9], taking document structure into account. In this paper, we are interested in extracting sentences that justify the citation context, thus helping readers make a relevance judgement about the cited document.

Sentence-related extraction technologies have been explored within the field of automatic text summarisation, although

[2] See www.elsevier.com.



Available online at www.sciencedirect.com

ScienceDirect

Vaccine 25 (2007) 7687–7695

www.elsevier.com/locate/vaccine

Further evaluation of higher potency vaccines for early protection of cattle against FMDV direct contact challenge

oyce, Scott M. Reid, Pip A. Hamblin,
Paton, Paul V. Barnett

oad Pirbright Surrey, GU24 0NF, United Kingdom

orm 24 July 2007; accepted 28 July 2007
4 August 2007

prior to severe direct contact challenge on protection from clinical
tigen payloads (single strength and 10-fold).
y reduce the number of clinically infected animals, and significantly
vaccinated group ($P < 0.05$). Although FMDV transmission occurred
tle held in indirect contact, no disease was induced in these animals.
antly reduce clinical disease and sub-clinical virus replication and
ity of transmission between animals and herds. To be most effective,
challenge significantly influences the number of animals becoming
aximum opportunity for protective immunity to develop.

strate the efficacy of emergency, high potency vaccines to protect from clinical and sub-clinical infection and to reduce transmission [1–8]. However, there is still a need to further substantiate emergency vaccination as a useful and additional method for control.

We recently published work on cattle, vaccinated 21 days prior to severe direct contact challenge, which indicated that higher payload vaccines conferred clinical protection and prevented, or reduced, sub-clinical infection and subsequent viral excretion, particularly during the early post exposure period [9,10]. Moreover, increasing antigen payload a further ten fold improved protection from sub-clinical infection and reduced the number of persistently infected animals. The work presented here makes a similar comparison, but with an earlier challenge at 10 days after immunisation. A pilot study was also included to assess transmission from vaccinated-challenged cattle to similarly vaccinated cattle.

25/03/2009 6:55:31 PM

Ref:Abstract:Citation-Sensitive Preview

S.J. Cox, C. Voyce, S. Parida, S.M. Reid, P.A. Hamblin, G. Hutchings (2006) Effect of emergency FMD vaccine antigen payload on protection, sub-clinical infection and persistence following direct contact challenge of cattle in *Vaccine*. 24 p.3184--3190.

Author Affiliation: :
S.J. Cox. Institute for Animal Health, Pirbright Laboratory, Ash Road, Woking, Surrey GU24 0NF, United Kingdom.

Abstract: :
Previous work, in sheep vaccinated with emergency foot-and-mouth disease (FMD) vaccine, indicated the benefit of increasing the antigen payload in inhibiting local virus replication and consequently persistence following an indirect aerosol challenge with a virus homologous to the vaccine strain. The work presented here investigates this possibility further using cattle and a more severe semi-heterologous direct contact challenge. The quantitative dynamics of virus replication and excretion in both vaccinated and non-vaccinated cattle following challenge are examined. Two experiments were carried out each involving 20 vaccinated and 5 non-vaccinated cattle. An O 1 Manisa vaccine (18 PD50) was used for the first, previously reported experiment [Cox SJ, Voyce C, Parida S, Reid SM, Hamblin PA, Paton DJ, et al. Protection against direct contact challenge following emergency FMD vaccination of cattle and the effect on virus excretion from the oropharynx. Vaccine 2005;23:1106-13]. The same vaccine was used for the second experiment described in this paper except the antigen payload was increased 10-fold per bovine dose, resulting in significantly higher FMD virus neutralising antibody titres prior to challenge. Twenty-one days post-vaccination the cattle received a 5-day direct contact challenge with FMD virus from five further non-vaccinated cattle infected 24 h earlier with O UKG 34/2001. All vaccinated cattle regardless of antigen payload were protected against clinical disease. Sub-clinical oropharyngeal infection was detected in animals from both experiments but the level of virus replication shortly after direct contact challenge was significantly reduced in vaccinated animals. Cattle immunised with the 10-fold antigen payload cleared the virus more readily and consequently at 28 days post-challenge fewer animals were persistently infected compared to the single strength vaccine. Following a severe challenge, the results from both experiments show that use of emergency vaccine can prevent or decrease local virus replication and thereby dramatically reduce the amount of virus released into the environment, particularly during the early post-exposure period. Additionally, increasing the antigen payload of the vaccine may reduce sub-clinical infection, leading to fewer persistently infected virus carrier animals.

Citation-Sensitive Preview: :

Introduction: It is well documented that **exposure** of **vaccinated** ruminants to FMD virus **infection** can protect against disease but still permit **sub-clinical** oropharyngeal **infection** and that such **infection** can persist.
Materials and methods: Briefly, **21 days prior** to **challenge,** 20 cattle were **vaccinated** with this increased antigen **payload** O1 Manisa vaccine, prepared from antigen concentrate, which is being held over liquid nitrogen as part of a new UK strategic reserve A further five animals remained unvaccinated as controls.
Results: The mean quantities of FMD **viral** RNA [log10(copiesml −1 .)] in probang samples found positive (non-zero) on quantitative RT-PCR, from both **vaccinated** and unvaccinated control cattle at different times after **contact** with the donor **cattle,** are presented in Fig. 2. During the **early period** between 4 and 10 **days** post-challenge exposure, the levels of oropharyngeal **viral** RNA detected in the unvaccinated cattle of both experiments was between 10 2 . and 10 3 . greater than that seen in the **vaccinated** cattle at the same time points.
Results: The differences in quantity of **viral** RNA detected between the **vaccinated** and unvaccinated cattle was analysed using ANOVA during this time **period** and at 7 **days** post-challenge exposure, significantly less (P less than ;0.05) **viral** RNA was detected from the **vaccinated** animals in both experiments.

Figure 2: An preview pop-up containing an automatically generated summary. Double clicking the citation in Adobe Acrobat activates this pop-up which allows the researcher to select and use the text in the summary.

with a differing application perspective from ours. The detection of sentences with citation information has been studied in the context of patent information by [6] (for Japanese documents). In earlier work, such sentences were used for the automatic compilation of survey articles [7].

Summarisation through the use of co-citation information has also been explored. An overview is provided in [2]; in that work, the authors analyse the similarity between citation contexts for a particular document. Similarly, [11] presented an approach for summarising an academic document based on the citations of that document found the literature at large. Citation contexts have also been classified according to the rhetorical role that they play in the document that is currently being read [13]. Given this classification, cited documents are summarised by showing sentences that support the specific rhetorical role. These approaches are similar to ours in that they exploit the use of citations between documents as linkages. However, except for [13], these approaches can only produce summaries that are independent of the reading context, while our approach deliberately tailors a summary to the citation context that has just been read. Our work differs from [13], in that we tailor the summary to the content of the citation context and not the rhetorical role.

There are a number of products that also cater to users in the academic domain. Knowledge extraction systems focusing on biomedical text, such as *BioMed Experts*,[3] *Illumin8*,[4] and *ConceptWeb*,[5] can automatically extract facts or highlight linkages (for example, by using clustering approaches) between researchers and published research. Often these tools provide improved search interfaces by suggesting ways in which to expand a query. These tools are complementary to browsing-oriented tools like CSIBS. Once documents are found, researchers can use CSIBS to navigate further through the space of document citations.

Reference managers, such as *Sente*[6] and *JabRef*,[7] can automate many tedious processes, such as automatically downloading cited documents and populating bibliographic databases. While these reduce the workload involved in obtaining referenced material, they do not help reduce the information overload on the researcher, who may still have to read each document in order to determine its relevance. Furthermore, this functionality runs the risk of placing researchers in a situation where they do not remember why they downloaded the document: at the time of downloading, the citation context, which provides that information, is no longer readily available.

3. OUR USER STUDY

To determine what readers of scientific literature require of cited documents, we conducted a user requirements analysis. We focused on users of biomedical literature because

[3] www.biomedexperts.com
[4] www.illumin8.com
[5] www.wikiprofessional.org
[6] www.thirdstreetsoftware.com
[7] jabref.sourceforge.net

we had access to an extensive corpus of material in this domain. However, our study and the questions we asked of participants were not specific to this domain.

We contacted thirty-six researchers in biomedical research, asking them to answer an on-line questionnaire. We explicitly stated the aims of the questionnaire by means of an initial brief. The first questions were multiple-choice questions based on general usage. For example, we asked about the reasons why participants used scientific literature and provided potential answers (e.g., "To learn about a new topic", "To update your knowledge on a particular topic", "Other: please specify"). We also asked them about the frequency of their literature browsing activity. Twenty-four participants completed that part of the questionnaire. Two thirds reported that they browsed through academic literature at least once a week.

The remainder of the questionnaire was aimed at uncovering researchers' information needs and tasks while carrying out a literature search. Our aim was also to find out what made the task difficult and how it could be supported with an automatic tool. Scenario-based questions asked participants to consider their day-to-day activities, with answers provided in free text. The questionnaire was limited to ten questions. Eighteen participants completed the entire questionnaire. Their responses were collated and analysed for commonalities, bringing to the fore those issues that were salient amongst the participants. The findings from the questionnaire can be summarised as follows:[8]

1. There were two main types of information needs, met respectively by *meta-level* and *content-oriented* information. These correspond to the tasks of *appraising the cited document* (to make a value-judgement about it) and *determining whether the citation was justified*, respectively. Both tasks in turn serve to help make a relevance judgement about the cited document.

2. The main difficulty participants reported was that of finding text that justified the citation context.

3. Searches within the document to find passages of text that justify the citation can be automated, guided by key words from the citation context, provided that the results are understandable by the user.

As a result of these findings, we chose to build a tool that meets the two types of information needs we uncovered, ultimately helping with the relevance judgement task. Our tool provides a summary of the cited document that contains both meta-data and content-oriented information, sourced from within the cited document. The former supports the article appraisal task, while the latter supports the task of citation justification; the information required is provided by automatic text summarisation techniques.

4. CITATION-SENSITIVE IN-BROWSER SUMMARISATION

In this section, we provide an overview of CSIBS, focusing on how the summary is generated and delivered on multiple presentation media.

4.1 The CSIBS System

Our system architecture is presented in Figure 3. CSIBS produces an preview-annotated version of a published academic document, rendered either as an Adobe Portable Document Format (PDF) document or a HyperText Markup Language (HTML) web page. The annotated document is a modified version of the original with CSIBS summaries inserted in the appropriate citation contexts. With the annotated document, the reader can activate the summary with mouse interactions in the appropriate browser.

We now describe the process involved in generating the annotated document using data retrieved from databases maintained by Elsevier. CSIBS is first provided with an identifier for the document that the reader has requested. This information is easily obtained from portals housing scientific data, such as ScienceDirect. The system retrieves the XML version of this document.[9] All citation instances are identified in the XML; these are then extracted and matched with the full references at the end of the document. Meta-data for each of the cited documents is retrieved from the *EMBASE* web service.[10] This provides the Elsevier-specific Publication Item Identifier (PII) number, amongst other things, which is used to retrieve the XML representation of each *cited* document from the Elsevier XML Repository.[11]

Each citation instance, its meta-data and the XML for the corresponding cited document is then passed to the summarisation engine, which computes the summary. The resulting summary contains a subset of the meta-data for appraising the cited document, and a citation-sensitive preview consisting of sentences extracted from the cited document to justify the citation: the meta-data corresponds to a generic view of the cited document and is independent of the citation context, while the citation-sensitive preview will differ depending on how the cited document is referred to in the citation context. Finally, the summary is inserted into the appropriate place in the annotated version of the published article.

To facilitate the researcher's literature browsing activities, CSIBS can be deployed as a web service attached to an existing publications portal (provided that appropriate permissions to databases are in place), so that the portal can provide the preview-annotated documents.

4.2 Meta-Data Summaries

CSIBS currently has functionality required to extract and present a variety of structured data, as follows:

- The full reference: This is useful in providing readers with information about the publication. It allows informal judgements based on the impact factor of the hosting journal publication, an attribute that many participants in the user study highlighted as being useful.

- The abstract: This is retrieved and presented as a generic overview: that is, a summary that is independent of the reading context.

[8] A paper describing the full details of the user requirements analysis has been submitted for review.

[9] While the functionality of CSIBS can be ported to other document collections, the current system assumes the availability of XML data.

[10] See `www.embase.com`.

[11] http://labs-repo.elsevier.com/repo

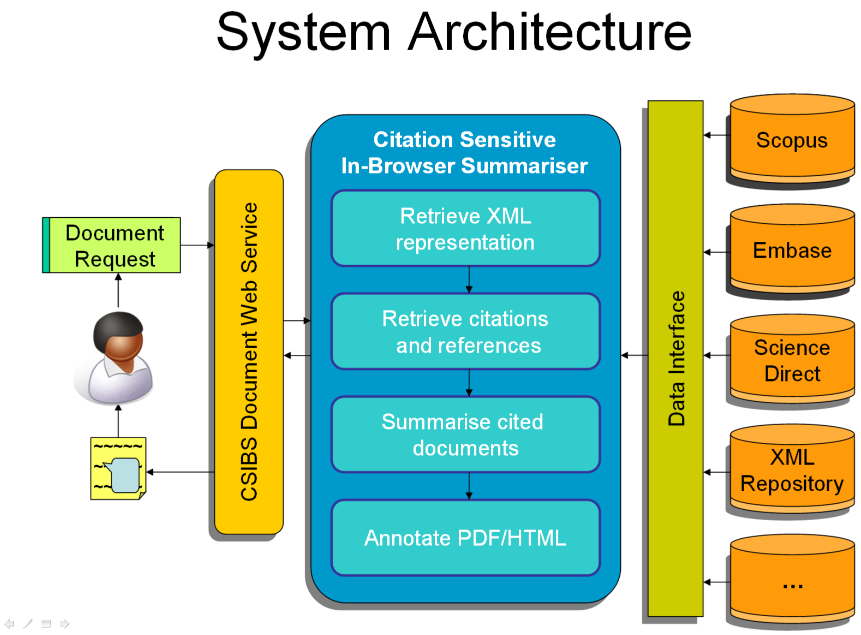

System Architecture
Document Request
CSIBS Document Web Service
Citation Sensitive In-Browser Summariser
Retrieve XML representation
Retrieve citations and references
Summarise cited documents
Annotate PDF/HTML
Data Interface
Scopus
Embase
Science Direct
XML Repository
...

Figure 3: A system architecture diagram. CSIBS produces *preview-annotated* PDF documents and HTML web pages. These are delivered as a web service.

- The structure of the cited document: This was also reported by participants in the user study to provide a useful snapshot of the document. CSIBS includes functionality to extract this for inclusion in the generated summary for readers to skim the cited document 'at a glance'.
- Author Information: CSIBS can include data to help the reader establish a level of trust in the citation, primarily focusing on information about the authors' affiliation and the number of related citations in the research area.
- The citation count for the cited document: This is extracted to help users appraise the citation.

Our development of CSIBS has been influenced significantly by the user requirements analysis briefly presented in Section 3. We have concentrated on structured data types for the appraisal task. Figure 4 presents a sample summary showing the possible meta-data as displayed in the HTML version of an preview-annotated ScienceDirect webpage. Figure 5 shows the same summary, this time with the abstract portion 'expanded' and the rest of the summary 'hidden'. Hiding and expanding all components of the summary is a user-interface feature, implemented in Javascript.

4.3 Citation-Sensitive Previews

4.3.1 Finding Justifying Sentences

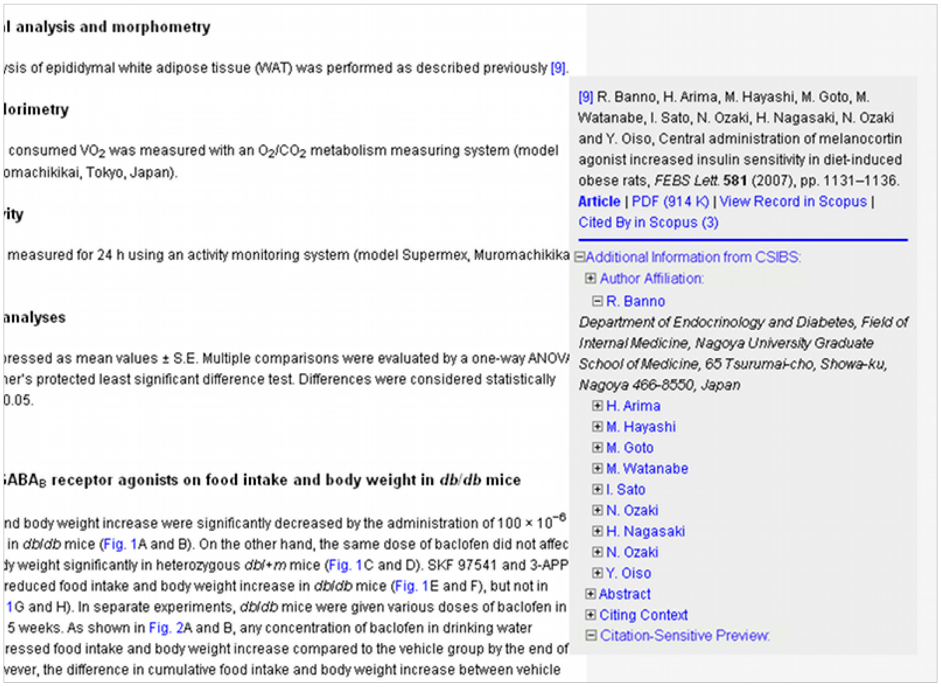
The module that generates the citation-sensitive preview requires, as input, the citation context and the set of sentences belonging to the cited document. The latter is easily obtained by extracting the text from the XML representation of the document and performing sentence segmentation. As output, the module produces a list of sentences for the citation justification task. These are extracted based on locating matching words (not including stop words). We cap the number of extracted sentences at a predefined limit, n, currently set to four.

To select sentences, we represent the citation context and each sentence in the cited document as a vector, where each dimension represents a word in the vocabulary, and the value for that position in the vector is the term frequency. We use vector space methods [12] to find the best n sentences that match the citation context, based on the cosine similarity metric. The attractiveness of this approach lies in its simplicity, speed and scalability. This is a crucial design element for a web-based service, for which interaction times must be reasonably fast.

We have chosen to use simple vector space approaches also because the user can clearly see how results were obtained. We found this to be a desirable trait for an automated system in our user requirements analysis (see Finding 3 in Section 3). Precise word matches which are highlighted allowing the reader to understand why particular sentences were extracted. Relevant section headings of the cited document were also provided as a means of grouping extracted sentences to improve the organisation of the citation-sensitive summary. Figure 6 presents two citation-sensitive previews side by side, each generated for the same cited document but triggered by different citation contexts. These are taken from a preview-annotated ScienceDirect webpage created by CSIBS. Since the citation contexts are different, the resulting citation-sensitive preview is different.

4.3.2 Narrow Information Trails

Exact token matching may, at times, fail to find the best



l analysis and morphometry

ysis of epididymal white adipose tissue (WAT) was performed as described previously [9].

lorimetry

consumed VO_2 was measured with an O_2/CO_2 metabolism measuring system (model
omachikikai, Tokyo, Japan).

ity

measured for 24 h using an activity monitoring system (model Supermex, Muromachikika

analyses

ressed as mean values ± S.E. Multiple comparisons were evaluated by a one-way ANOV
her's protected least significant difference test. Differences were considered statistically
0.05.

ABA$_B$ receptor agonists on food intake and body weight in *db*/*db* mice

nd body weight increase were significantly decreased by the administration of 100 × 10^{-6}
in *db*/*db* mice (Fig. 1A and B). On the other hand, the same dose of baclofen did not affec
dy weight significantly in heterozygous *db*/+*m* mice (Fig. 1C and D). SKF 97541 and 3-APP
reduced food intake and body weight increase in *db*/*db* mice (Fig. 1E and F), but not in
1G and H). In separate experiments, *db*/*db* mice were given various doses of baclofen in
5 weeks. As shown in Fig. 2A and B, any concentration of baclofen in drinking water
ressed food intake and body weight increase compared to the vehicle group by the end of
vever, the difference in cumulative food intake and body weight increase between vehicle

[9] R. Banno, H. Arima, M. Hayashi, M. Goto, M. Watanabe, I. Sato, N. Ozaki, H. Nagasaki, N. Ozaki and Y. Oiso, Central administration of melanocortin agonist increased insulin sensitivity in diet-induced obese rats, *FEBS Lett.* **581** (2007), pp. 1131–1136.
Article | PDF (914 K) | View Record in Scopus | Cited By in Scopus (3)

⊟Additional Information from CSIBS:
⊞ Author Affiliation:
⊟ R. Banno
Department of Endocrinology and Diabetes, Field of Internal Medicine, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550, Japan
⊞ H. Arima
⊞ M. Hayashi
⊞ M. Goto
⊞ M. Watanabe
⊞ I. Sato
⊞ N. Ozaki
⊞ H. Nagasaki
⊞ N. Ozaki
⊞ Y. Oiso
⊞ Abstract
⊞ Citing Context
⊟ Citation-Sensitive Preview:

Figure 4: A sample pop-up summary rendered on ScienceDirect page.

justification sentences for citations. The optimal justification sentences might not share words with the citation context, due to differences arising from surface-level string variation, resulting from, for example, the use of synonyms or morphological variants of a word.

Ideally, the CSIBS system should record a match between two tokens where surface-level string variations exist and yet the semantics of the two words are sufficiently close. However, given the limited screen real-estate of the interface, and the need to provide understandable results, care must be taken in terms of allowing the system to relax its word matches to account for these string differences. CSIBS should not relax its sentence selection constraints only to burden the reader with poorly matching sentences, as this will merely exacerbate any existing information overload problems.

One possibility is to use manually constructed lexical resources that specify sets of related terms and synonyms. In the context of Life Sciences scientific literature, numerous biomedical ontologies exist that specify the relationships between terms. Methods for using such knowledge resources have been explored in text summarisation: the use of thesaural resources for finding topically related terms was explored in [4], and resources such as Wikipedia have been mined as a similar resource of related terms [8]. While these methods can, at times, improve the sentence matching, this comes at a cost. False positives — irrelevant sentences that match loosely related words — can make the resulting summary unusable.

In response to this, we propose a method, *Narrow Information Trails*, that errs on the side of caution. With this method, matches with associated words are permitted only under very strict circumstances. We simply compile a list of permissible matches to related words using two hard filters:

1. we only consider words we know to be reasonable summary words for the cited document; and

2. we restrict this expanded set of words to those that have a string similarity above a certain threshold.

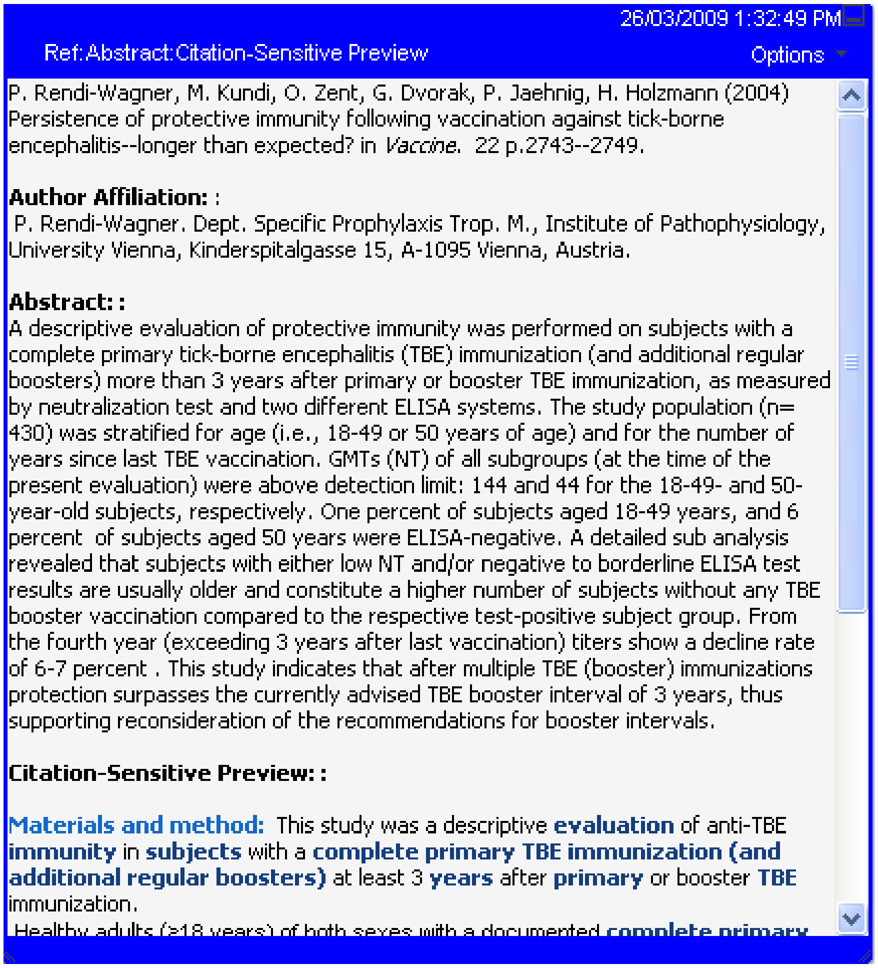
To construct the first filter, we investigated the use of *co-citations* as a means of identifying good summary words for a cited document. Co-citations are citation contexts in other documents in the literature that refer to the same cited document. This approach stems from the observation that the ways in which other authors refer to a document may provide an excellent and precise lexical resource of related words. This observation was also made by [11] and [6], who consider co-citation contexts as potential generic summaries (thus not tailored to the reading context).

The second filter is implemented as a function based on Levenshtein String Distance, as follows:

$$\text{Match}(s_1, s_2) \quad = \quad 1 - \frac{\delta(s_1, s_2)}{\max(|s_1|, |s_2|)}$$

where s_1 and s_2 are two strings of length greater than 4, and δ is the Levenshtein String Distance function. The threshold we currently use for this function is 0.5.

We take each word in the original citation context and compare it to words returned by the first filter. Comparisons are made using the string distance approach. Words that are above the threshold are used in addition to the words from the original citation context to select sentences. We weight the original citation context words higher using a fixed factor, currently hardcoded to be three times the



P. Rendi-Wagner, M. Kundi, O. Zent, G. Dvorak, P. Jaehnig, H. Holzmann (2004) Persistence of protective immunity following vaccination against tick-borne encephalitis--longer than expected? in *Vaccine*. 22 p.2743--2749.

Author Affiliation: :
P. Rendi-Wagner. Dept. Specific Prophylaxis Trop. M., Institute of Pathophysiology, University Vienna, Kinderspitalgasse 15, A-1095 Vienna, Austria.

Abstract: :
A descriptive evaluation of protective immunity was performed on subjects with a complete primary tick-borne encephalitis (TBE) immunization (and additional regular boosters) more than 3 years after primary or booster TBE immunization, as measured by neutralization test and two different ELISA systems. The study population (n= 430) was stratified for age (i.e., 18-49 or 50 years of age) and for the number of years since last TBE vaccination. GMTs (NT) of all subgroups (at the time of the present evaluation) were above detection limit: 144 and 44 for the 18-49- and 50-year-old subjects, respectively. One percent of subjects aged 18-49 years, and 6 percent of subjects aged 50 years were ELISA-negative. A detailed sub analysis revealed that subjects with either low NT and/or negative to borderline ELISA test results are usually older and constitute a higher number of subjects without any TBE booster vaccination compared to the respective test-positive subject group. From the fourth year (exceeding 3 years after last vaccination) titers show a decline rate of 6-7 percent . This study indicates that after multiple TBE (booster) immunizations protection surpasses the currently advised TBE booster interval of 3 years, thus supporting reconsideration of the recommendations for booster intervals.

Citation-Sensitive Preview: :

Materials and method: This study was a descriptive **evaluation** of anti-TBE **immunity** in **subjects** with a **complete primary TBE immunization (and additional regular boosters)** at least 3 **years** after **primary** or booster **TBE** immunization.
Healthy adults (≥18 years) of both sexes with a documented **complete primary**

Figure 5: A sample pop-up summary with the abstract of the cited document.

original term frequency, to maintain a focus on the current reading context.

Our work has similarities with the approach in [10] which also looks at co-citation contexts as a source of paraphrase information, including synonyms. In contrast to that work, which uses deep text analysis approaches, our method, based on string difference, is targeted at variations in terms due to morphological differences. By narrowing equivalence strictly to string distance, we are able to match additional words that must *look* similar, thus making the fuzzy matching criteria transparent to the users. Additionally, we do this without the use of heuristic stemmers or manually-constructed dictionaries, which may be limited in use only to particular domains.

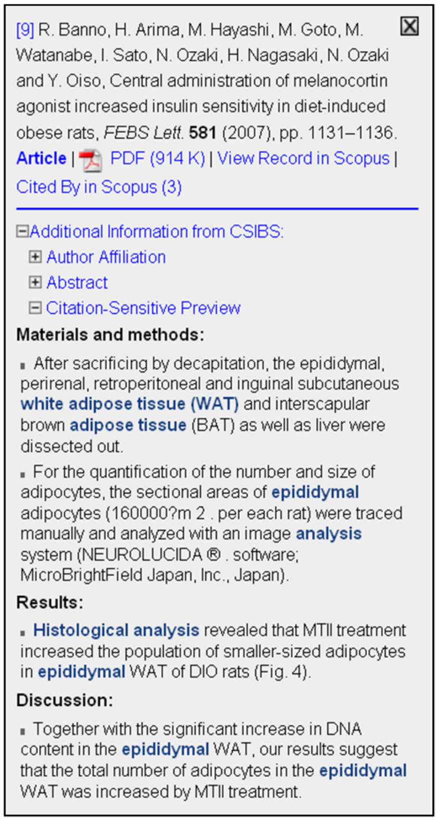
Figure 7 presents an example of some co-citation contexts and a diagram of the corresponding narrow information trails. The citation context shown in this example is the same as that shown in Figure 2. We use the *Scopus* database,[12] which provides documents that are co-citing. The co-citation contexts from these extra documents are extracted and used to construct the narrow information trails shown in the figure. As can be seen, this can have a beneficial impact on the citation-sensitive preview. For this citation, an additional sentence is found that now outranks previously identified justification sentences. This was found based on a matching of the additional word *Chlorobium*, which relates to the word *chlorosomes*.

4.4 The Scalability of CSIBS

The ability to generate summaries depends on the required information being available. CSIBS makes use of existing databases maintained by a publisher; the benefit of using such resources is that they provide fast response-times to complex database queries, such as finding all documents that co-cite a document. The time to generate an preview-annotated document is currently dominated by the time it takes to retrieve the cited documents. Currently, our demonstrator retrieves these documents one at a time using a research API to the repositories. In a real production system, the data would be co-hosted with the CSIBS web-service and queries to databases would be sent in parallel, thus effectively removing network lag time.

Once data has been retrieved from the databases, the time to create a summary is bounded by the number of citation contexts, c, and the largest number of sentences in a cited

[12] www.scopus.com



[9] R. Banno, H. Arima, M. Hayashi, M. Goto, M. Watanabe, I. Sato, N. Ozaki, H. Nagasaki, N. Ozaki and Y. Oiso, Central administration of melanocortin agonist increased insulin sensitivity in diet-induced obese rats, *FEBS Lett.* **581** (2007), pp. 1131–1136.
Article | PDF (914 K) | View Record in Scopus | Cited By in Scopus (3)

⊟Additional Information from CSIBS:
⊞ Author Affiliation
⊞ Abstract
⊟ Citation-Sensitive Preview

Materials and methods:

- After sacrificing by decapitation, the epididymal, perirenal, retroperitoneal and inguinal subcutaneous **white adipose tissue (WAT)** and interscapular brown **adipose tissue** (BAT) as well as liver were dissected out.
- For the quantification of the number and size of adipocytes, the sectional areas of **epididymal** adipocytes (160000?m 2 . per each rat) were traced manually and analyzed with an image **analysis** system (NEUROLUCIDA ® . software; MicroBrightField Japan, Inc., Japan).

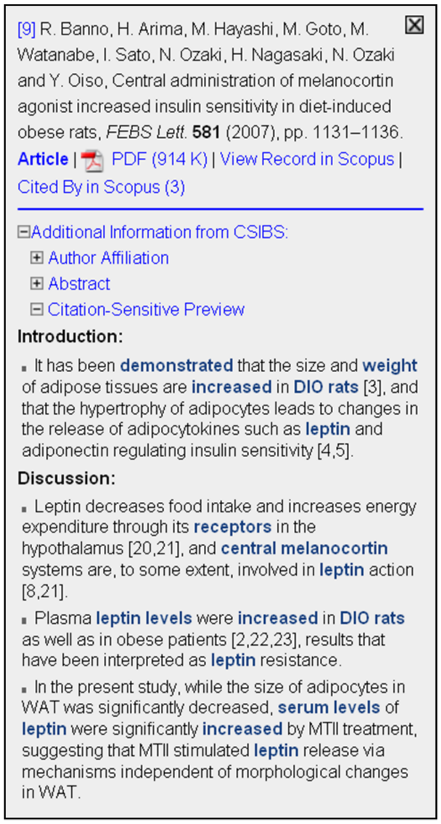
Results:

- **Histological analysis** revealed that MTII treatment increased the population of smaller-sized adipocytes in **epididymal** WAT of DIO rats (Fig. 4).

Discussion:

- Together with the significant increase in DNA content in the **epididymal** WAT, our results suggest that the total number of adipocytes in the **epididymal** WAT was increased by MTII treatment.

Histological analysis of epididymal white adipose tissue (WAT) as performed as described previously [9].



[9] R. Banno, H. Arima, M. Hayashi, M. Goto, M. Watanabe, I. Sato, N. Ozaki, H. Nagasaki, N. Ozaki and Y. Oiso, Central administration of melanocortin agonist increased insulin sensitivity in diet-induced obese rats, *FEBS Lett.* **581** (2007), pp. 1131–1136.
Article | PDF (914 K) | View Record in Scopus | Cited By in Scopus (3)

⊟Additional Information from CSIBS:
⊞ Author Affiliation
⊞ Abstract
⊟ Citation-Sensitive Preview

Introduction:

- It has been **demonstrated** that the size and **weight** of adipose tissues are **increased** in **DIO rats** [3], and that the hypertrophy of adipocytes leads to changes in the release of adipocytokines such as **leptin** and adiponectin regulating insulin sensitivity [4,5].

Discussion:

- Leptin decreases food intake and increases energy expenditure through its **receptors** in the hypothalamus [20,21], and **central melanocortin** systems are, to some extent, involved in **leptin** action [8,21].
- Plasma **leptin levels** were **increased** in **DIO rats** as well as in obese patients [2,22,23], results that have been interpreted as **leptin** resistance.
- In the present study, while the size of adipocytes in WAT was significantly decreased, **serum levels** of **leptin** were significantly **increased** by MTII treatment, suggesting that MTII stimulated **leptin** release via mechanisms independent of morphological changes in WAT.

… increased serum leptin levels in DIO rats while the body weight was decreased by the injection [9].

Figure 6: Although both pop-ups display citation-sensitive previews for the same cited document, the contents of the are not the same because the citation context is different.

document, s, and has complexity $O(c \times s)$. This is almost negligible (in the order of seconds) in comparison to the data retrieval, which, once co-hosted on the same server, should no longer be an impediment.

5. THE UTILITY OF CSIBS

5.1 Evaluation Overview

A qualitative evaluation was performed to gauge whether the generated summaries provide additional utility beyond a full reference when judging the relevance of cited documents. Three biomedical researchers, who had taken part in our initial user requirements analysis, participated in our evaluation. Each participant was asked to read a short passage containing an annotated citation and to decide if the citation was relevant given the citing context (Question 1). They were also asked whether they would follow up on the citation (Question 2). For both questions, participants had to indicate the strength of their answer on a 7-point Likert scale.

The task was performed under three conditions, depending on the pair of preview-annotated PDF documents presented to the participant. These were displayed within Adobe's Acrobat Reader. In each case, the enhancement consisted of either the full reference, the abstract of the cited document, or a citation-sensitive preview of that document. Each of the three participants was randomly assigned to a condition and shown 9 annotated documents, each being an article from the journal *FEBS Letters*.[13]

In the first condition, for each trial, the participant was provided with a PDF document annotated only with the full reference and asked to read a set passage within the document. The participant answered the two questions for the specified embedded citation. The participant was then shown a variant of the same document annotated with the abstracts and asked to perform the same task. The second condition was the same as the first except that the second annotated document contained a citation-sensitive preview. The third condition was the same as the second, except that

[13] www.febsletters.org

The Citation Context:

- The green sulfur bacterium has chlorosomes containing multi-layered and single -layered tubular structures of self-aggregated bacteriochlorophyll (BChl) [citation]

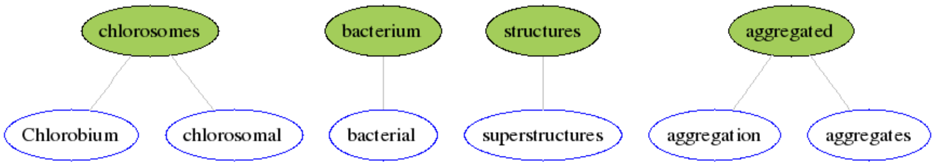
The Cited Document:

- Marie Ãstergaard Pedersena, Jarl Underhauga, Jens Dittmera, Mette Millerb and Niels Chr. Nielsen. (2008) The three-dimensional structure of CsmA: A small antenna protein from the green sulfur bacterium Chlorobium tepidum. In *FEBS Letters*. 582. p2869-2874.

Co-Citation Contexts:

- the exact nature of BChl superstructures remain elusive and these have been controversially discussed in the literature
- cryo-TEM both on green bacterial chlorosomes from Chlorobium tepidum
- Either up-running stacks or down-running stacks can form extended domains, which have been imaged by TEM
- Very recent high-resolution images confirm that chlorosomes are flattened cigar-shaped nanostructures with dimensions of 20 50 200 nm
- Although the self-aggregation of BChl c, d, and e is well known, as mentioned above, the exact structure of chlorosomal aggregates has been a matter of controversial discussion for a long time
- The lamellar nature of the BChl aggregates was confirmed in a very recent electron cryomicroscopy study

Information Trails:



Additional Summary Sentence (ranked #1):

[10] The chlorosomes of the model green sulfur bacterium **Chlorobium** tepidum have been extensively characterized biochemically.

Figure 7: Information trails using co-citation contexts. This figure presents the original citation context, the full reference for the cited document, co-citations contexts, the information trails, and the additional matched sentence. The diagram presents words of the original citation context in filled nodes, with words from the trails in unfilled nodes. In the newly matched sentence, the new word, *Chlorobium*, is matched and thus presented in bold.

an additional task was imposed: the participant was also asked to indicate which of the three extracted sentences in the citation-sensitive preview were useful in performing the tasks, thus providing utility feedback for each sentence. In all conditions, participants were free to ask questions. Written and verbal comments were captured throughout the evaluation.

5.2 Results

We measured the change in the strength of the responses, or the self-reported confidence, to each of the two questions. For Question 1, across the first two conditions, there was a mean change in self-reported confidence of 1.22 in favour of the abstract; and there was a mean change of 2.22 in favour of the citation-sensitive preview. Although it is not possible to draw any firm conclusions given the small sample size, it is encouraging that the citation-sensitive preview led to a stronger self-reported confidence. For Question 2, judgement confidence increased with a mean change in self-reported confidence of 1.33 in favour of abstracts, and 1.44 in favour of citation-sensitive previews. The third condition also revealed a mean utility for selected sentences of 63 percent.

The final participant, familiar with *FEBS Letters*, made some particularly interesting observations regarding the selected sentences and the structure of the cited document. Specifically, useful sentences tended to be found deeper in the cited document, for example in the methods sections. Additionally, the participant remarked that, for corporate laboratories where each document downloaded from a proprietary repository incurs a fee, the generated summaries would make it easier to decide whether to download a document by providing a much needed citation-related preview. She also mentioned that, in the absence of sufficient information, she would err on the side of not downloading the document.

The results suggest that providing summaries, whether they are abstracts or citation-sensitive previews, allows the reader to make relevance judgements with more confidence, as compared to being presented with just a full reference. The citation-sensitive preview was useful if it offered more specific details: that is, when sentences were extracted from sections other than the introduction. We believe this is because the greater level of detail complements the genericity of the abstract.

6. FUTURE WORK

The work described in this report presents CSIBS as a proof-of-concept. In our pilot study, the tool was judged to be useful. Although CSIBS is only a research prototype, care was taken with the system design and engineering to easily optimise CSIBS in the future to scale it up for deployment as a live web service.

From a research perspective, we intend to investigate methods for tuning the summarisation performance via our narrow information trails approach. We also intend to engage further with users directly, both to obtain additional insights into information needs, and to evaluate the current system in greater depth. For example, a number of system parameters are currently predefined, such as the citation-sensitive preview sentence cap and the fixed selection of meta-data information. However, these parameters can be set experimentally with user studies, or potentially left to the user to specify dynamically.

Finally, we have demonstrated the techniques we have developed using the literature of one scientific discipline, that of biomedicine. We aim to also apply the techniques to other disciplines; we expect the transfer to other scientific domains to be straightforward, but the extent to which our approach can be used more broadly remains to be seen.

7. CONCLUSION

In this paper, we described a new research tool aimed at supporting researchers as they acquire knowledge through the perusal of scientific literature. The Citation-Sensitive In-Browser Summariser (CSIBS) was designed and developed by taking into account real tasks and information needs of academic researchers, particularly biomedical researchers. CSIBS helps readers determine whether or not a cited document is worth reading further, addressing issues of trustworthiness and citation justification. This is done by providing readers with documents annotated with summaries. We evaluated CSIBS by presenting the resulting annotated documents to three users, who indicated that the extracted summaries were useful for the relevance judgements. The approach naturally complements a document's own abstract, given that generated summaries can present detailed information from the body of the cited document that is more relevant to the specific citation context.

8. ACKNOWLEDGMENTS

We would like to thank the participants in the initial user study and the three trial participants for the qualitative evaluation for their time and insightful comments. We would also like to thank Elsevier for permitting access to their data via the Elsevier Grand Challenge. Finally, we would like to thank Julien Blondeau, Michael Muthukrishna and Ilya Anisimoff, for their efforts in performing engineering tasks and HCI studies.

9. REFERENCES

[1] N. J. Belkin. Design principles for electronic textual resources: Investigating users and uses of scholarly information. In *Current Issues in Computational Linguistics: In Honour of Donald Walker.Kluwer*, pages 1–18. Kluwer, 1994.

[2] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, and D. Radev. Blind men and elephants: What do citation summaries tell us about a research article? *J. Am. Soc. Inf. Sci. Technol.*, 59(1):51–62, 2008.

[3] H. Li, W.-C. Lee, A. Sivasubramaniam, and C. Giles. Workload analysis for scientific literature digital libraries. *International Journal on Digital Libraries*, 9(2):139–149, November 2008.

[4] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[5] Y. Liu, K. Bai, P. Mitra, and C. L. Giles. Tableseer: automatic table metadata extraction and searching in digital libraries. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 91–100, New York, NY, USA, 2007. ACM.

[6] H. Nanba, N. Anzen, and M. Okumura. Automatic extraction of citation information in japanese patent applications. *International Journal on Digital Libraries*, 9(2):151–161, November 2008.

[7] H. Nanba, N. Kando, and M. Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *The 11th SIG Classification Research Workshop, Classification for User Support and Learning, 2000.11, in Chicago, USA*, pages 117–134. The American Society for Information Science (ASIS), 2000.

[8] V. Nastase. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 763–772, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[9] T. Nguyen and M.-Y. Kan. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326, 2007.

[10] N. I. Preslav, A. S. Schwartz, and M. A. Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Workshop on Search and Discovery in Bioinformatics*, 2004.

[11] V. Qazvinian and D. R. Radev. Scientific paper summarization using citation summary networks. In *The 22nd International Conference on Computational Linguistics (COLING 2008)*, Machester, UK, August 2008.

[12] G. Salton and M. J. McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.

[13] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computional Linguistics*, 28(4):409–445, 2002.

[14] S. Wan and C. Paris. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Paper*, Columbus, Ohio, June 2008.

[15] J. Zhao, M.-Y. Kan, and Y. L. Theng. Math information retrieval: user requirements and prototype implementation. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 187–196, New York, NY, USA, 2008. ACM.