CONTENT ENRICHMENT THROUGH DYNAMIC ANNOTATION

George R S Weir
Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

George Lepouras, Costas Vassilakis
Department of Informatics and Telecommunications, University of Athens, Athens, Greece

Keywords: Dynamic annotation, Web content enrichment

Abstract: This paper describes a technique for interceding between users and the information that they browse. This facility, that we term 'dynamic annotation', affords a means of editing Web page content 'on-the-fly' between the source Web server and the requesting client. Thereby, we have a generic way of modifying the content displayed to local users by addition, removal or reorganising any information sourced from the World-Wide Web, whether this derives from local or remote pages. For some time, we have been exploring the scope for this device and we believe that it affords many potential worthwhile applications. Here, we describe two varieties of use. The first variety focuses on support for individual users in two contexts (second-language support and second language learning). The second variety of use focuses on support for groups of users. These differing applications have a common goal which is content enrichment of the materials placed before the user. Dynamic annotation provides a potent and flexible means to this end.

1 INTRODUCTION

Our research over the past five years suggests that a desirable characteristic in many application areas is the ability to intercept information requested by a Web user and modify its content on-the-fly before it arrives on the user's display. Such activity has no sinister intent, but affords an effective means of supplementing or filtering Web content in a 'just in time' fashion so that the user's information exposure is better honed to the needs or wishes of that user. The technology required to affect such a system builds upon a common component in network architectures, viz., a caching Web proxy server.

In a Web context, many organisations require that users browse the Web in a controlled and monitored fashion. Often, this entails the use of a Web proxy facility such as 'squid' (http://www.squid-cache.org/). These systems sit between a user's client machine and remote Web servers. The local Web client software is configured to relay all Web requests to the proxy server. In turn, this proxy forwards the request to the appropriate remote site. Since the remote Web server receives the request from the proxy machine, its response is directed to the address of the proxy server. Upon receipt of the requested document, the proxy server determines which internal client should receive this response and relays the document to the appropriate host.

Such behaviour provides a number of security advantages to the local network. In the first place, local users cannot access external Web sites except via the proxy server. Thereby, the proxy provides a central point at which a record of all Web browsing can be maintained. This proxy log is a useful device for monitoring the behaviour of Web users by recording who is browsing and what information they are viewing. A second benefit of Web proxies is that they often allow administrators to impose restrictions on the local user population by curtailing access to specific sites. A third proxy benefit is the fact that such servers often provide a caching facility. This reduces duplicate requests for remote documents, in line with the expiry times specified on retrieved Web pages, and thereby reduces data

R S Weir G., Lepouras G. and Vassilakis C. (2004).
CONTENT ENRICHMENT THROUGH DYNAMIC ANNOTATION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 152-157
DOI: 10.5220/0002643301520157

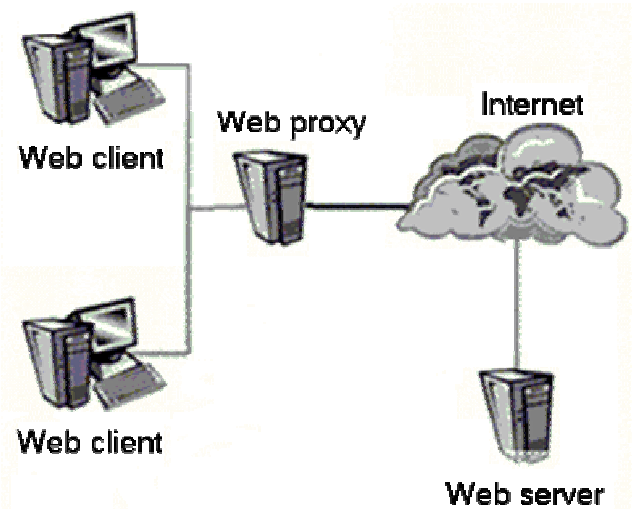

Figure 1: Web proxy in network configuration.

traffic to and from the network. Finally, use of a Web proxy affords a measure of protection for local client machines. Since external servers only see the proxy address, the addressing scheme and number of machines behind the proxy can be concealed from external view. The usual network position of the Web proxy is illustrated in Figure 1.

2 DYNAMIC ANNOTATION

To this context of proxy-based Web browsing we bring the possibility of intercepting incoming Web pages, scanning, and optionally modifying their content. This facility we term 'dynamic annotation'. 'Dynamic' reflects the fact that content can be modified on-the-fly between the user's request and delivery of the selected Web page. 'Annotation' denotes a principal use that we have found for this mechanism.

Note that our use of annotation is more generic than other Web annotation approaches (e.g., Kahan et al, 2001 and Ovsiamnikov et al, 1999). Commonly, such systems apply annotation as a commenting facility available to those viewing Web-based information. In such contexts, users are able to add their comments by annotating the information that they view. Contributed comments are then visible to other users. For example, Annotea (op. cit.) supports 'shared annotations (comments, notes, explanations, or other types of external remarks that can be attached to any Web document or a selected part of the document)'. Significantly, Annotea does not modify the original Web document but stored comments externally in an annotation server. A metadata facility for describing annotations and a pointer mechanism for locating the annotations in the annotated document are the basis for unifying the document and annotation display for the end user.

While a manual facility is included in the third of our annotation applications (described below), the feature that differentiates our approach to annotation is that the system itself adds 'extra' information to the Web pages requested by the user. This is closer to the Magpie approach described by Dzbor et al (2003). In the following, we outline the operation of our annotation facility and describe its use in several example applications.

2.1 Annotation mechanism

In order to achieve the interception and modification of Web content we have employed a Web server that supports proxying. The Apache server (cf. http://www.apache.org) fits this bill and has the added benefit that it allows the use of customised dynamically loadable modules. This convenience permitted the development of an add-on module that filters the content of retrieved Web pages. The filtering process relies upon a parsing mechanism that, in turn, applies a dictionary of terms. These are the terms upon which some modification is required.

Our initial versions of the annotation facility used a static dictionary with pre-defined terms. The later version supports a dynamic dictionary, allowing the terms of interest to be changed during the operation of the annotation system. This feature provides a basis for adapting the Web content modification strategy during its operation. Work on this adaptive mechanism, which employs feedback from the user's interaction, is currently in progress. The architecture for the dynamic annotation facility is shown in Figure 2.

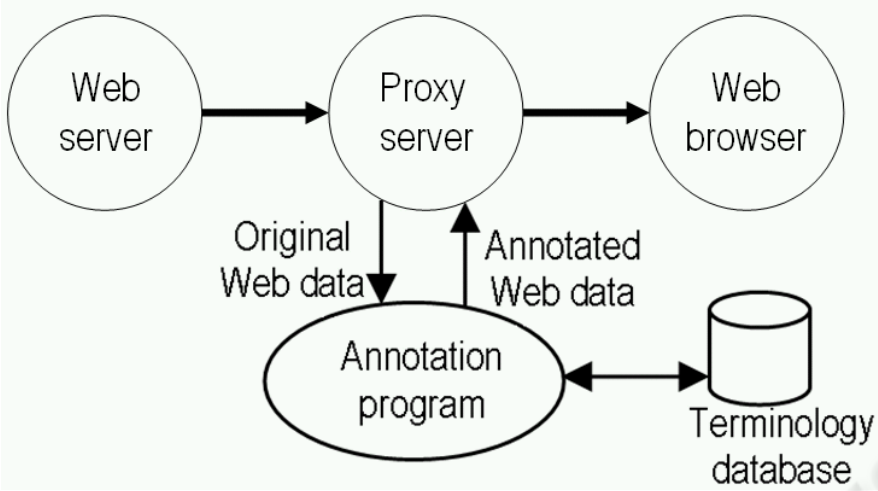

Figure 2: Dynamic annotation mechanism.

2.2 Interaction support

We have developed applications using dynamic annotation for two different contexts in which the user's native language is a significant factor. The first case is the context of second language support the second is the context of second language learning.

Second language support is an approach that we have developed to the 'second language problem' in human-computer interaction (Weir et al., 1996; Lepouras & Weir, 1999). This is the setting in which non-native speakers of English are obliged to operate with English-based computerised information systems. A common consequence in this scenario is difficulty in user interaction. For such individuals, language introduces particular difficulties and misunderstanding in the use of such interactive systems.

One might expect that fully language-localised software products or information systems would solve all problems of language misunderstanding for the end user. Experience suggests that this is not the case. Instead of full remedy, the localisation strategy adds or replaces existing problems with new ones. We have characterised several sources of such problems (Lepouras & Weir, 1999), including:

1. the absence of standardised terminology;
2. mixed language environments, and
3. cross-language confusion.

At first sight, Web information systems seem an unlikely setting for the second language problem. Web pages deliver information, so language comprehension will be an issue but the relevance of this to user interaction is not obvious. The difficulty is that reading information from Web pages is a major component of the interaction context. Web users do more than read delivered information. The Web is an interactive system in which the user must negotiate a route through selected information resources. This means that the user's ability to comprehend the visible labels on Web links will affect their choice of successive resources.

The user selects links based upon presumptions of their relevance and interest value. Where hyperlinks rely upon textual description, the user's interpretation of the text plays a dominant role in determining interaction. From this we conclude that language comprehension significantly affects the success of user interaction in the Web context. In consequence, user's whose first language differs from that used in the Web pages are more likely to experience degraded interaction. This is the feature of the second language problem.

Our work on second language support has focused on examples in Greek and Chinese. Although these languages require font sets that complicate their use in conjunction with English language displays, we are able to combine local language support (e.g., Chinese text) with the original language (i.e., English).

In aiming to support the second-language user we could aim for full translation of Web page content, e.g., English to Chinese, but translation is not our principal concern. In addition, experience suggests that targeted use of local language support can enhance user interaction by improving comprehension whilst minimising additional risk of language-based dissonance. We have explored a variety of mechanisms for supplementing Web information with native language annotations. One approach adds selective translations in the form of pop-ups, to pre-selected English terms. An example of this technique is shown in Figure 3.

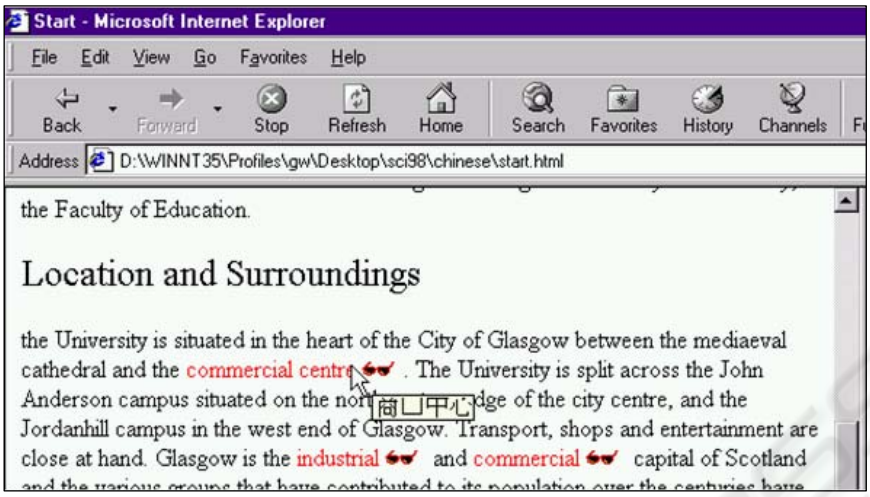

Figure 3: Chinese annotation on English Web content.

In this example, English terms that may not be readily understood by the native Chinese speaker are annotated with Chinese translations or explanations. This is achieved by searching incoming Web page content, locating terms that require elucidation and modifying the Web page content to include the Chinese addendum. In this instance, we have changed the HTML code to include an icon and generate pop-ups with suitable Chinese language content. The target English expressions are also altered so that they appear in red. This enables the Chinese user to identify readily those terms for which supplements are available.

Other annotation techniques include the use of frames to add a second source of commentary on the English original. This technique appeals to those users who prefer that the original content is left unaltered. Whatever the specific annotation facility, each has the aim of content enrichment.

2.3 Language learning support

A related but distinct area of application for dynamic annotation is a setting in which users wishing to learn English may use Web browsing as a basis for enriching their familiarity with English language. We have explored the use of Web-based content enrichment as a means of adding explanations in the native language of the user (Weir & Lepouras, 2001). This provides a native language support that does not seek to smooth user interaction but rather aims to facilitate better understanding of English language and texts expressed therein.

A particular strength in this use of dynamic annotation is that language learners may freely explore Web sites without constraint. The support facility is tuned to add annotation in a variety of forms to specific sets of English terminology. This context is considerably richer in varieties of support since the variability of linguistic needs may be extensive.

An area that merits greater exploration is the application of criteria to determine the deployment of native language annotation. While our work in this aspect is still at an exploratory stage, we have identified several plausible factors to serve as drivers for such dynamic support:

1. Learner selected experience level
2. Frequency table
3. Fog index
4. Meta information

The simplest approach is learner driven. This provides a range of support levels and allows learners to vary the degree of annotation applied. Other techniques, such as frequency analysis of English terms and expressions, are potential supplements to user selection. This information would enable the selective application of native language support, based upon the likely exposure of the learner to such terms. In addition, a Fog Index may be determined for each paragraph, in order to gauge the likely need for clarification. Such metrics as the Gunning fog index (Gunning & Kallan, 1994) provide a heuristic measure of text readability based upon the average number of words in each sentence and the number of polysyllabic words in the given sample of t Although our purpose is not to modify original English content, the fog metric is a viable technique for guiding annotations. Furthermore, such insight may focus attention on paragraphs to which automated summarisation techniques may be

appropriate. Finally, a rating system may be introduced to signify the language complexity for specific Web pages. Either this could be provided by enlightened authors as 'meta' content to their pages or generated dynamically and added to cached versions of incoming pages. Through dynamic annotation techniques driven by metrics for English language complexity, there is scope to provide a virtual 'English Assistant' that will ensure that English content is appropriately enriched to support user learning.

2.4 Group support

Our application of dynamic annotation has been extended to a third context: group support (cf. Lepouras et al, 2003). Here, we have a Web-based conferencing facility that supports the management and dissemination of domain experience in an organisation's operational environment. In this context, workers often share valuable experience and knowledge but this is usually exchanged in an informal rather than formal manner. This tacit or implicit knowledge is often lost as it resides in email and memos that afford no easy method of retrieval, classification, or cross reference.

The content enriched system that we have prototyped allows for dynamic classification of local knowledge in a systematic manner that eases its retrieval from the knowledge database. In operation, the system requires that individuals login and configure a profile of user preferences. Thereafter, the user can register for topic-specific discussions, exchange messages with other system users, create direct links to favourite or interesting messages, annotate messages and rate them according to the validity or importance of their content (Figure 4).

Dynamic annotation makes an appearance through automatic recognition of key terms and automatic enhancement of messages via auto-generated links to other messages with content of relevance to topics in which the user has expressed an interest (Figure 5). In concept, this shares some aspects of the 'ComMentor' system developed at Stanford University (Roscheisen et al., 1994). One major difference is that our system is active in automatically cross-referencing content and in its use of term equivalence in support of user searches.

At the heart of the system's classification component is an explicit taxonomy of the organisation. This taxonomy includes lists of equivalent terms whose nodes can be interconnected; thereby the user may search for a term in messages and retrieve a message even though the exact term does not appear in the search pattern.

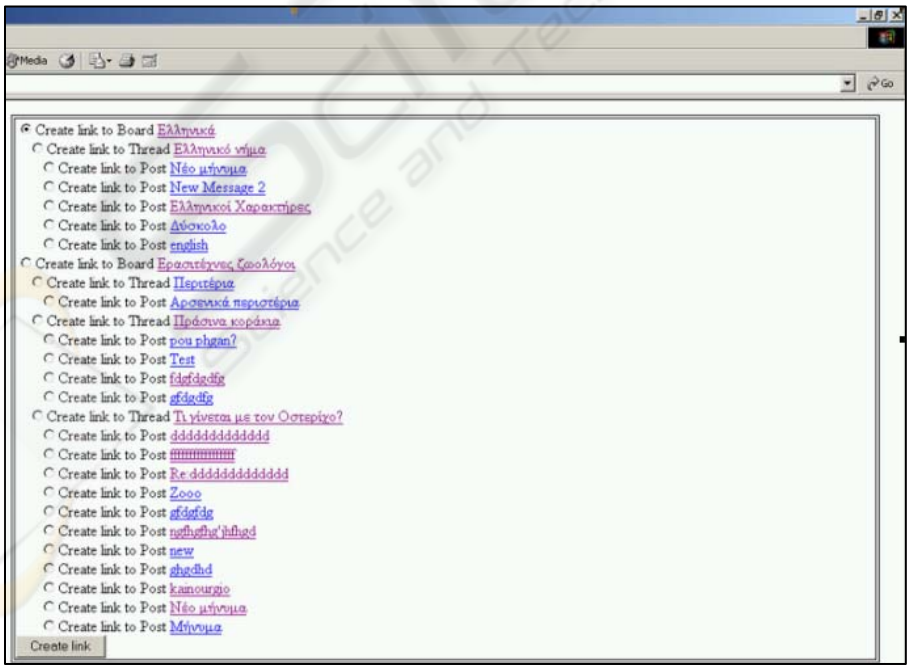

Figure 4: Link creation dialogue.

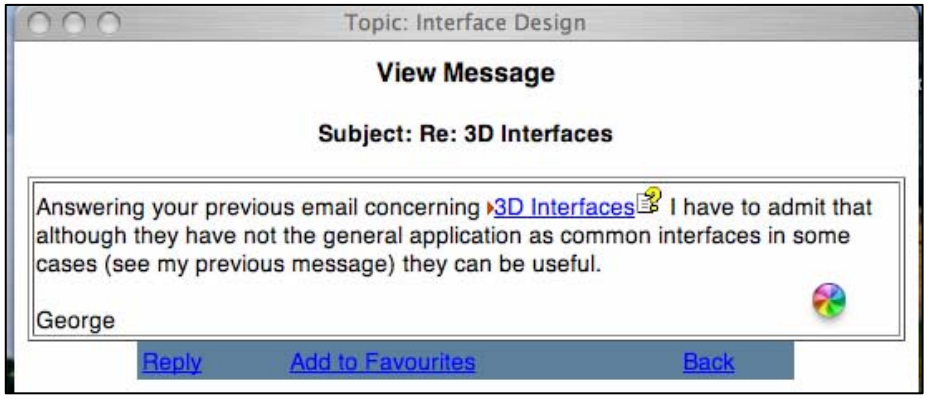

Figure 5: Message with annotated keywords.

The taxonomy and the general discussion categories are maintained through the system administration screens. Through these screens the administrator can define new general discussion categories, connect them to the taxonomy and add key-terms to describe the topics. In its present form, this application has extensible content enrichment but is not yet adaptive.

3 GENERAL APPLICATION

Our purpose in this paper is to detail the benefits of dynamic annotation as a means of content enrichment and to indicate the general applicability of this approach wherever the delivery of Web-sourced information can be enhanced to better suit the needs or interests of the individual user.

The variety of application areas extends from supporting the interaction of individuals for whom English is a second language, through to annotation in the context of group interaction. Between these extremes, dynamic annotation affords support for those seeking to use Web resources as a basis for learning English, and as a means of building upon otherwise implicit institutional knowledge.

Next on our research agenda is the addition of feedback from user interaction with content and annotations as a basis for building adaptivity into the mechanism for dynamic annotation. This will allow us to explore ways of allowing user behaviour to direct the mode and content of annotation.

REFERENCES

Dzbor, M., Domingue, J. & Motta, E., 2003. Magpie – towards a semantic web browser. In *Proceedings of the 2nd International Semantic Web Conference 2003 (ISWC 2003)*, Florida, USA

Gunning R. and Kallan R. A., 1994. *How to take the fog out of business writing*, Dartnell.

Kahan, J., Koivunen, M. R., Prud'Hommeaux, E. & Swick, R. R., 2001. Annotea: An Open RDF Infrastructure for Shared Web Annotations. In *Proc. of the 10th World Wide Web Conference (WWW10)*, Hong Kong.

Lepouras, G., Vassilakis, C, Weir, G. R. S., 2003. A System to Support Dissemination of Knowledge and Sharing of Experiences in the Working Environment, International Journal of Continuing Engineering Education and Life-Long Learning, Volume 13, Nos. 3/4, 2003, 248-257.

Lepouras, G. and Weir, G.R.S., 1999. It's not Greek to me: terminology and the second language problem', *ACM SIGCHI Bulletin*, vol.31 No.2, (April), 17-24, ACM Press.

Lepouras, G. and Weir, G.R.S., 2003. Subtitled interaction: Complementary support as an alternative to localization, *International Journal of Human Computer Studies*, Volume 59, Issue 6, December, 941-957.

Ovsiannikov, I.A., Arbib, M.A., and Mcneill, T.H., Annotation Technology. *International Journal of Human-Computer Studies*, 1999. 50(4): p. 329-362.

Roscheisen, M., Mogensen, C. & Winograd, T., 1994. Shared Web annotations as a platform for third-party value added information providers: Architecture, protocols, and usage examples, *Technical Report CSDTR/DLTR*, Stanford University.

Weir G.R.S. and Lepouras G, 2001. English Assistant: A Support Strategy for On-Line Second Language Learning, *IEEE International Conference on Advanced Learning Technologies*, IEEE Press.

Weir, G.R.S., Lepouras G. & Sakelleridis, L., 1996. Second language help for Windows applications, in M. A. Sasse, R. J. Cunningham and R. L. Winder (Eds.) *People and Computers XI, Proceedings of HCI'96*, Springer, London, 129-138.