ISSN: 2321-7782 (Online)
Volume 2, Issue 5, May 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Literature Survey on Annotation based Web Systems by Using Query Records from Web Databases

Sonali T. Kadam[1]
Department of Computer Engineering
Bhivarabai Sawant Institute of Technology and Research
University of Pune
India

Sanchika Bajpai[2]
Department of Computer Engineering
Bhivarabai Sawant Institute of Technology and Research
University of Pune
India

Abstract: Annotations are mostly useful mechanisms that support an amount of advantageous document management applications like highlighting a text, inserting additional notes, third-party commentary, design rationale, data filtering, and semantic labelling of manuscript. The ubiquity of web database content influences the need for web annotation systems that are lightweight, efficient, transparent, platform-independent, user friendly and scalable. Constructing such a system using open and regular web infrastructures facilitates extensive applicability and deployment. The paper describes our literature survey experiences with client, browser-based web annotation systems and proxy server based annotation system details. We have observed that web communication systems are missing elements like recent web infrastructure that create any implementation of annotation systems unsatisfactory. Technical hitches with web-based text information retrieval (IR) systems include unauthentic matches, manually intensive document sifting, and the lack of communication or coordination between users. A prototype implementation, and the collaborative infrastructure enabled by Annotate! System has improved the transmission of ideas in the search community. This paper discusses fundamentals of current web infrastructure, potential changes and annotation based web communication systems to the web design that might make the proposal of annotation systems more complete.

Keywords: Web Systems, Annotation, Web Database, Query Extraction, Web Communication.

I. Introduction

Web annotation has ambitions to turn the web into several medium. It enables a page that is originally constructed by a single author who defines the initial content of the document, to be modifiable by any interested contributor. In fact, more than authoring the most pervasive activity around documents after reading is annotation followed by collaboration and authoring [1]. Although Web annotation has been an active area of research since the genesis of the World Wide Web, progress has been slow mainly due to the constraints of the Web infrastructure lack of annotation standards, and the slow adoption of the emerging standards in Web browsers, e.g., XML linking technologies.

Despite speedy progress in hardware computing platforms in modern years and equally rapid improvements in the popularity of the Web as a distributed hypermedia publishing system, information retrieval (IR) remains a thorny problem equally in the Internet and in organizational intranets. Popular web-based full text search (WFTS) engines like an Excite, Lycos, and Alta Vista use a hypermedia interface on the front-end Query Interface, but on the back end the information Retrieval Interface is just an array of hyperlinks pointing back to source documents. From the user's viewpoint, the chief failing of an ad-hoc Web IR system is the lack of data and metadata clues in the search, retrieval, and document browsing interfaces.

The three interfaces are the Query Interface - access point of a search engine, the Retrieval Interface - a usual result from a full text search engine is a group of hyperlinks to base documents, and the Document Interface- the result from clicking on a link in the Retrieval Set is typically a single document browsing session. There is no metadata (data about the documents) available at the document layer; author and timestamp information must be explicitly represented in the document body. The restricted set of clues coupled with the semantically weak GOTO mechanism [2] coupling the Retrieval and Document layers makes for continued inefficiency in users' interactions with a web search system. The user must navigate the Retrieval Set, one document at a time, with very limited clues e.g. a summary paragraph, the document's title, and the engine's confidence score to indicate if that particular document might be important to solving the original problem. The Retrieval Interface poses a formidable challenge to the user must browse, laboriously, one document ata time from the Retrieval Set with only limited data and metadata clues as signposts that might point to a document actually relevant for the problem at hand. And, since web based search engines typically do not circumvent the statelessness of the Web's Hypertext Transfer Protocol (HTTP) we have no memory is kept between search sessions [3] for a given user, and experiences for better or worse with specific queries cannot be communicated between users.

Software tool, Annotate![4]is presented, to address these challenges by adding a collaborative dimension to Web full text search. In Annotate! two data sets, declared in XML, which are at the core of Annotate!: discussion data, a composite of documents and user annotations and session data which captures user timings at the many interface layers.

Synchronous Web annotations systems are exceptional. A one or two of examples can be found online and in the research literature. These include IBM Markup, GroupWeb, and MemoChat[5] and there is no commercial product in this classification yet. IBM Markup system and MemoChat lack even the most basic meta-data available in simple text Chat programs, such as name tagged messages. Text annotations in the IBM's Markup system can appear anywhere on top of the document, but users cannot specify to specific parts of the document they are associated.

New Web technologies such as DOM and DHTML are demonstrating useful for building more interesting, advanced and user-friendly annotation clients. For instance, it is possible to insert new objects in a HTML document at specified positions and change the layout of a table with dynamism. At present only a handful of systems have explored the potential of these advances of the Web infrastructure. Some of the greatest known systems to take advantage of these developments are: IBM's Markup System [6], ThirdVoice [7] (defunct), Yawas [8],and WebAnn [9]. Though, the exploration is hardly complete. None of the modern systems has considered the use of DOM to support the incremental augmentation of HTML documents with open functionality. Extensibility can increase adaptability to users' changing requirements and new technologies as well as rapid prototyping and testing new features. Currently, there are two main approaches to achieve feature extensibility on the Web, by the use of a programmable proxy and using a proprietary browser.

This paper provides the contribution in evolution of annotation, different annotation based web communication systems and annotation systems, current and future services in annotation and collaboration literature review. Further paper will also introduce comparison of currently existed annotation systems. The some sections discuss literature survey of annotation based web communication system such as in building proxy-based, server-based web annotation systems, client-based annotation systems and browser-based Web annotation systems. It has been our experience from different annotations surveying that annotation systems are constrained both in capability and efficiency by the limitations of current web infrastructure.

We find that the intermediary approach offers a reasonable, uniform structure for extending web client capability - externally to the browser. Yet there is little support to date for extending capabilities within popular browsers in the same principled way, due to security mechanisms and contrary browser designs.

II. Literature Survey

This section summarizes previous and ongoing recent projects that subject to support the annotation of web documents. Only rare attempt are made to be exhaustive, as the goal is to compare and contrast annotation based web systems with efforts presented in [5], [10], [11].Existing annotation systems vary in terms of implementation approach and functionality for the particular purpose system was designed. In essence, they all change some aspects of the Web infrastructure e.g., browser, content, web protocol with transparency to the user [10]. The approaches that these projects adopt can be broadly classified in terms of the locus of augmentation, the place where the annotations and/or annotation capabilities are incorporated into the Web document displayed by the browser. This is done via an intermediary agent [12] that is located wherever in the path between the Web server and the Web browser: at origin i.e., Web server, in proxy server which can be external or local to the client PC, or at arrival i.e., Web browser. Intermediary agents trigger the annotation process by intercepting page requests, contents of Web pages, or events (e.g., page loading).

The ability to annotate web documents provides a mechanism that can be the basis of a number of useful document management applications. Annotations allow third-parties to interactively and incrementally augment web documents. An annotation system supports the creation and retrieval of annotations, and composes personalized "virtual documents" from the authored document and associated annotations. Intermediary agents trigger the annotation process by intercepting page requests, contents of Web pages, or events (e.g., page loading). The systems that introduce web annotation function without changing web content, browsers or servers. E.g. Strand/GrAnT[13] and Net Notions [14].

Abstract annotation system architecture is successfully implemented using client or server-side Internet frameworks as shown in [5]. The architecture consists of different main components like interceptors, annotation repository services (AReS) and composers, with the annotation delivery and composition styles being personalized based on a user model. The implementations point to missing elements (e.g current web infrastructure, and potential changes to the web architecture) in the recent web infrastructure that might make implementation of annotation systems more perfect and complete. In the next section we will discuss the different types of annotation based web communication systems.

In [15], a document-independent framework concept allows contextualized synchronous discussions to take place in the form of text chat was introduced. In order for this framework to work, the specific document application must be adapted to communicate with an external IRC (Internet Relay Chat) client.NCSA's Mosaic project [NCSA] and ComMentor[16] change web browsers to augment them with annotation capabilities, and are the examples of a customized infrastructure approach to building annotation systems.

III. Annotation Fundamentals

A. *Annotation:* We have defined web annotations as “Online annotations associated with web resources such as web pages, with which users can add, update or delete information such as highlighting, commentary, link making, reading records from a web page without modifying the page itself”.

Many purposes of making annotations have been identified like text annotation; IT based annotation, web annotation, Audio- video annotation, Image annotation, JAVA annotation, XPS annotation as explained in [1], [4], [5], [10], [12].In [5] authors reviewed and suggested web annotation system’s architecture, that includes new technologies such as the Document Object Model (DOM) level 2 will be desirable to design high-quality annotation systems. “Conceptual architecture of the individual mode of WATs (Web Annotation Tools)” and annotation example, which is demonstrated in Fig.1 and Fig.2 respectively.

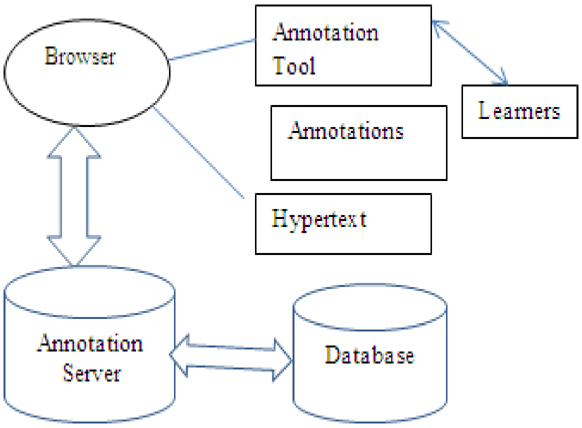
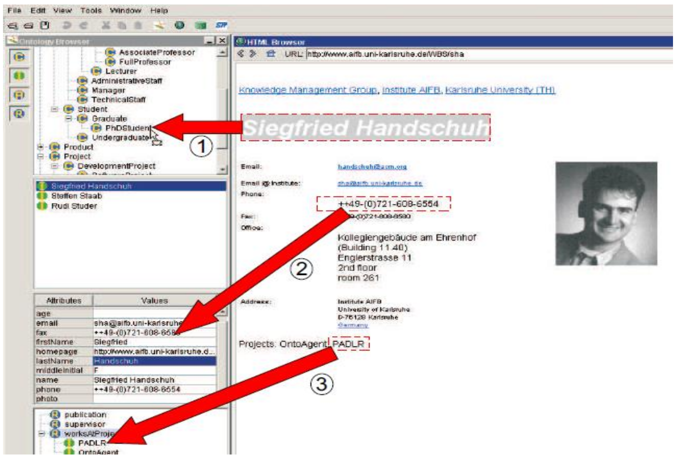


Fig.1 Architecture of web annotation



Attributes	Values
age	
email	sha@aifb.uni-karlsruhe
fax	++49-(0)721-608-658
firstName	Siegfried
homepage	http://www.aifb.uni-karlsruhe.d...
lastName	Handschuh
middleInitial	F
name	Siegfried Handschuh
phone	++49-(0)721-608-6554
photo	

Knowledge Management Group, Institute AIFB, Karlsruhe University (TH)

Siegfried Handschuh

Email: handschuh@acm.org
Email @ Institute: sha@aifb.uni-karlsruhe.de
Phone: ++49-(0)721-608-6554
Fax: ++49-(0)721-608-6580
Office: Kollegiengebäude am Ehrenhof
(Building 11.40)
Englerstrasse 11
2nd floor
room 261
Address: Institute AIFB
University of Karlsruhe
D-76128 Karlsruhe
Germany

Projects: OntoAgent, PADLR

Fig 2.Annotation Example

Implementation of collaborative information retrieval in Annotate! is layered on top of the Excite core search engine with a series of small server-side Perl modules. Excite is a convenient choice because the distribution which is no-cost, licensed software available for many operating systems platforms includes the source code for the interface libraries. Such libraries control the look and the application on the front-end and the search results record returned from the search engine. Annotate! is independent, though, of the core search engine and can be used in conjunction with experimental search algorithms [4] or enhanced user-interface frameworks [4]. The first major advantage of collaborative information system is an improved and dynamic growing set of data and metadata clues available to the user in the various interface layers. Using the Retrieval metadata signposts should assist the user in minimizing wasted browsing time at the Document layer but this must be tested empirically. The Retrieval icons which rise from annotations are powerful in many ways such as they do not require user effort since they are bound to the interface changes via automated alteration rules, they reflect immediacy (the most recent annotations), assist navigation into the Document layer, and in the aggregate enable a social filter at the Query layer.

B. *Annotation Based Web Systems:* Here we are going to discuss the some existing web system in following paragraphs.

1. *Server-based Web annotation systems:* There are only few known systems since this category approach requires the main documents to be pre-processed in advance to have the essential hooks for annotations, and therefore, such systems could not be generalized to work for random HTML documents on the Web. CoNoteand Virtual Notes are the examples of these systems.

2. *Proxy-Based Web Annotation Systems*: Proxy-based methods store and merge annotations by using a proxy server. A user fetches documents they are interested in through a specific URL that acts as an intermediary between the user and the webpage. The proxy servers add a user interface and merge existing annotations with the page before displaying itto the user. Proxy-based implementations are gorgeous since they are less responsive to variations in browsers.

Perhaps, the biggest advantage of the proxy approach is its ability to operate at the HTTP stream level. This presents a great deal of flexibility for document augmentation. Consider the following scenario which can benefit from proxy intermediation: Dynamically generated page in form submissions. Suppose that the search result page of the submission is dynamically generated, the submit button should be disabled right after the document is submitted the first time, and the result page should be the same for all the assembly browsers.

The proxy can replace the form handler with a customized handler to guard the submit button and instruct the proxy to retain the result page and dispatch the same page to all the browsers Another strong point of the proxy-based approach is that it can be easily extended to provide the browser with access to local resources, such as the client’s file system, since there are less security restrictions for an application than for an applet.

On the other hand, a major disadvantage of the proxy approach is the requirement of the proxy itself. The deployment requires installation and browser configuration, and if the operating environment already requires that the browser be configured with proxy indirection, then the proxy of the annotation system may not be accessible at all. The well-known examples are: Critlink and Annotator.

Proxy-based interception is justly easy and flexible when compared to client-side interception. Proxies, particularly those built on top of Java web servers allow programmable access both to the outgoing request object and the incoming document content. The former is used to implement request interception, and the latter can be used to implement page interception. One limitation of proxy-based interceptors, which is really a limitation of the HTTP protocol, is that the proxy cannot distinguish between request for a document, and requests for sub-objects e.g. embedded images of the document.

3. *Browser-based Web annotation systems:* Increasingly more implementations move the mediator agent to the browser, occupy more user-friendly annotation placement. Selection user interface based on highlighting, and support annotations of arbitrary web documents. For such systems, users often have to install a small software program that adds a user interface for creating annotations to their Web browser and then can make and view annotations as they read HTML documents. However, by extending a particular browser the systems can often take advantage of non-standard features of that browser. Browser-based Web annotation systems are: ComMentor [20] is the first known web annotation system.

IV. Comparison of Annotation Systems

Table I describes different existing annotation system tools with their different features and table II shows the comparison of the existing annotation based web communication systems in terms of their augmentation approach, support availability for extensibility, and native code dependency.

TABLE I COMPARISON OF WEB ANNOTATION SYSTEMS

Annotation system	Private notes	Private group notes	Public notes	Notification	Highligh ting	Forma tted text	Notes	Technology	Open Source
Firefox	Yes	No	No	No	No	No	This is provided by"Description" and "tags" fields of bookmarks.	Bookmarks	Yes
A.nnotate	Yes	Yes	No	Yes	Yes	No	tate PDF, ODF, .doc, .docx, images, and web pages can be annotated(but limited free version are there)	Snapshots	No
Chatterati	No	No	Yes	No	No	Yes	Presently available as a Google Chrome extension allows users to have a discussion, vote, comments.	Chrome extension	No
Crocodoc	Yes	Yes	No	No	Yes	No	Requires Adobe Flash and Can annotate PDF, .doc, images, web pages.	Snapshots	No
Delicious	Yes	No	Yes	No	No	No	per user 1000 character limit per page	Bookmarklet	No
Diigo	Yes	Yes	Yes	Yes	Yes	No	Public annotations are only allowed for established users.	Toolbar	No
Keeppy	Yes	Yes	No	Yes	Yes	No	Allows web annotationand stores the note on the cloud server.	Toolbar	No
Org-mode	Yes	No	No	Yes	No	Yes	requires technical knowledge to set up; not as user-friendly, non-Latin characters allowed in notes but tags are not permitted	Text Editor	Yes
rbutr	No	No	Yes	Yes	No	Yes	it allows public comments and tags to links.	Toolbar	No
Reddit	Yes	Yes	Yes	Yes	No	Yes	It is mainly intended for new and interesting links. Voting , links ranked by #votes and age;	Toolbar	Yes

TABLE II COMPARISON OF EXISTING ANNOTATION BASED WEB COMMUNICATION SYSTEMS

Year	System	Augmentation Approach					Extensibility		Page Detection	
		Browser-Based			Proxy-Based	Server Based				
		Applet	activeX Plug-in	Custom Browser			Runtime	Compile Time	Native Code	Nonnative Code
1994	ComMentor			P					P	
1995	CoNote					P				P
1996	Groupweb			P					P	
1997	Critlink				P					P
1998	Annotator				P				P	
1999	IBM Markup				P					P
1999	JotBot	P							P	
1999	Third Voice		P						P	
2000	Yawas		P						P	
2001	Multivalent			P				P	P	
2001	MemoChat		P							
2001	Annotea			P					P	
2001	iMarkup		P						P	

Note: P stands for present

V. Conclusion

From an ad-hoc application in which every user is quarantined from him or her in the time dimension i.e. no session memory and from the other users in proximate work groups in the time and distance dimension i.e.no group query memory and no group identifiers .Our goal is to move to a state-oriented application that provides useful clues as the user moves through search, retrieval, and web browsing. As the foregoing survey shows Web annotation systems designed for discussions are very restrictive in linking capabilities and inflexible in discourse. In addition, significant research efforts remain to be invested, in particular as tribute to the open socio-technical issues and interface designs for collaboration and communications using annotation. Investigation in Web annotation is extreme from complete. We have examined some of the research opportunities related to using Web annotation for dialogues. There are additional exciting critical problems which hamper Web annotation from becoming pervasive in one and all's daily Web activity. We found some of fields in which we can work and these are as follows.

Limited availability and invasive deployment: Annotation capabilities are neither universally available nor integrated seamlessly in all the browsers. In order to annotate a document on the Web, users have to install an annotation extension on their browser and/or change the browser settings.

Annotated object type restrictions: A Web page can have mixed object types, including non-HTML objects, such as ActiveX, scanned documents. The vast majority of systems only support text-only annotations that can be attached to HTML text portions of a Web document.

Annotation repositioning: Annotation systems often assume that the underlying annotated documents do not change. Positioning annotations robustly over changing document contents poses challenges other than technical. Security and annotation filtering: When annotations are stored in public annotation servers, the shared or private annotations should be protected from the reach of unauthorized users and the owners of web sites should be able to prevent impropriate annotations to appear on their sites. A best example of this is this problem faced Third Voice.

Acknowledgement

We express our thankfulness towards the professors of BSIOTR(w) and S.B.Patil COE for helpful discussions about this work.

References

1. Brush, A., Bargeron, D., Gupta, A., and Cadiz, J.J., "Robust Annotation Positioning in Digital Documents", in Proceedings of CHI 2001, pp. 285-292.
2. C.Watters, M. Conley, and C. Alexander. "The digital agora: Using technology for learning in the social sciences". Communications of the ACM, 41(1):50–57, January 1998.
3. J. W. Cooper and R. J. Byrd. "OBIWAN — a visual interface for prompted query refinement". In Digital Documents, volume 2, pages 277–285. 31st Hawaii International Conference on System Sciences, IEEE, January 1998.
4. Ginsburg, M., "Annotate! A Tool for Collaborative Information Retrieval", in Proceedings of WETICE'98, also at http://raven.stern.nyu.edu / papers/.
5. Ng S. T. Chong," Annotation-based Web Communications Systems: A Review", Technical report CS-3408,Oct-2003
6. Pacifici, G. and Youssef, A., "Synchronous Annotation of Shared HTML Documents", in Proceedings of the IASTED International Conference '99, 1999, pp. 275-279.
7. ThirdVoice.http://www.thirdvoice.com, 1999.
8. Denoue, L, and Vignollet, L., An Annotation Tool for Web browsers and its Applications to Information Retrieval, inProceedings of RIAO 2000, 6th Conference on Content based Multimedia Information Access, 2000
9. Brush, A, Bargeron, D., Grudin, J., Borning, A., and Gupta, A., Supporting interaction outside of class, in Proceedings of Computer Support for Collaborative Learning, 2002, pp. 425-434.
10. Vasuden, V. and Palmer, M., "On Web Annotations: Promises and Pitfalls of Current Web Infrastructure", in Proceedings of the 32nd Hawaii International Conference on Systems Sciences, 1999.
11. http://en.wikipedia.org/wiki/Web_annotation
12. Barrett, R. &Maglio, P. P., Intermediaries: New Places for Producing and Manipulating Web Content, in Proceedings of the 7th International World Wide Web Conference, 1997.
13. Schickler,M. et al., "Pan-Browser Support for Annotations and Other Meta-Information on the World Wide Web", in Proceedings of the Fifth International World Wide Web Conference, 1996
14. "NetNotions Product Details",http://www.sideware.com
15. Churchill, E., Trevor, J., Bly, S., Nelson, L., and Cubranic, D., Anchored Conversations: Chatting in the Context of a Document. In Proceedings of the 2000 ACM Conference on Human Factors in Computing Systems, 2000, pp. 454-461
16. Roscheisen,M. et al., "Content Ratings and Other Third-Party Value-Added Information Defining an Enabling Platform", D-Lib Magazine, August 1995, also at http://www.cnri.reston.va.us/home/dlib/august95/stanford/08 roscheisen.html

AUTHOR(S) PROFILE

Sonali T. Kadam completed her B.E. degree in Computer Engineering from VPCOE, University of Pune in 2009. She is pursuing M.E. in Computer Engineering from BSIOTR, University of Pune, Maharashtra, India. She has a engineering teaching experience of SBPCOE,UOP,India. Her research interest includes Knowledge and Data Engineering, Database Management System and Data Mining

Sanchika Bajpai is working as Assistant Professor at JSPM's BSIOTR, Pune, Maharashtra, India. She completed her B.Tech degree in IT from Chaudhary Charan Singh University in 2009 and M.Tech. in Computer Science from Amity University, Lucknow. Her research interest includes Software Testing and Database Management System.