User-centred systems of information retrieval in the digital era

Vishwa Mohan Vangari

Dept. of Library & Information Science, Osmania University, Hyderabad. India. drvvm321@yahoo.com

Abstract

A user-centred system of Information Retrieval is the order of the day. In the emerging digital information systems, there is a wide distance or gap between the information seeker or users of the information system and the information workers. When the digital information resources are available to distant information seeker, ensuring efficient and user-friendly or user-centred information retrieval systems is indispensable. In the Internet era, the online information systems endeavor to design and develop user-centred retrieval systems. But, these systems fail to satisfy most of the important norms of the information retrieval systems. Finally, they also prove to be user-unfriendly. The paper attempts to assess the feasibilities of designing and developing user-centred systems. It is quite clear that the nature, approach, level of knowledge, etc. of the users of information retrieval systems would vary very widely. In view of this the study attempts to find the diversity of user approach in a cross section of academics. And measure the amount of diversity and find the solutions for unification of the diverse user approach.

Keywords: Information searching, Literature searching, Retrieval Systems, User-centred indexing, User-centred Information Retrieval Systems, User-friendly Information.

Resumen

Los sistemas de recuperación de información (SRI) centrados en el usuario están de actualidad. En los sistemas emergentes de información digital existe una gran distancia entre los usuarios de los sistemas de información y los profesionales de la información. En un momento en que los recursos informativos digitales están disponibles para los usuarios que buscan información a distancia, resulta imprescindible asegurar la disponibilidad de SRI eficientes, amigables y centrados en el usuario. En la era de Internet, los sistemas de información en línea dirigen sus esfuerzos al diseño y desarrollo de sistemas de recuperación centrados en las necesidades del usuario. Sin embargo, estos sistemas no consiguen cumplir la mayoría de las principales normas de los SRI. Además, se ha demostrado que no resultan

amigables para el usuario. La comunicación trata de establecer las características del diseño y desarrollo de sistemas centrados en el usuario. Resulta evidente que la naturaleza, el enfoque, el nivel de conocimiento, etc. de los usuarios de los SRI varían en gran medida. Dada esta situación, este estudio intenta analizar la diversidad de aproximaciones de los usuarios académicos a la recuperación de la información en un cruce de disciplinas académicas, así como medir dichas diferencias de comportamiento y encontrar soluciones para unificar los diversos enfoques de los usuarios.

Palabras clave: Búsqueda bibliográfica, Búsqueda de información, Indización dirigida al usuario, Sistemas amigables de recuperación de la información, Sistemas de información centrados en el usuario.

1 Introduction

An Information Retrieval System (IRS) forms an integral and most crucial part of an Information System (IS). The earliest IRSs were not user-centred as they were manual systems which could not meet varied needs of different types of users. However, they endeavoured to be user-friendly by providing multiple access points and different cross-reference entries. As an attempt to transfom themselves into user-centred IRSs, the IRSs evolved themselves from those that used controlled vocabulary and fixed syntax to post-coordinate, keyword/uniterm (Taube, 1954) and permuted Indexing Systems (Sharp, 1965, Ferrandane, 1970, Austin, 1974). Even these systems suffered from many limitations. When the digital information resources are available to distant information seeker, ensuring efficient and user-friendly or user-centred information retrieval systems are indispensable as there is an increasing geographical/physical gap between the IS and the users, as a result there will be no scope for extension of personal help or assistance to the users in using the IRS. In view of this in the Digital Era, the machine retrieval systems have opened up a number of search options to the users enhancing the status of the IRSs from system-centred IRSs to User-centred IRSs.

Presently, there are very efficient IRSs. The Machine Retrieval Systems exhibit great retrieval capabilities, v. gr. speedy retrieval, omnibus retrieval, etc. But still, are they user-centred? The online information systems endeavor to design and develop user-centred retrieval systems. But, these systems fail to satisfy most of the important norms of the information retrieval systems. Finally, they also prove to be user-unfriendly. A study into this aspect would reveal that they are more efficient and less user-centred. Search Engines like Google, Altavista, Yahoo, etc. retrieve millions of sites/documents, in fractions of a second, of which very negligible percentage of them would be relevant to the users. How such Retrieval systems can be called as user-centered? They are neither user-centred nor user-friendly. What is the use of an IRS that causes considerable waste of time of a user by retrieving almost 99% or sometimes 100% irrelevant resources out of millions of resources? What is the use of high recall value with negligible relevance and precision rates? How much time would it take for a user to check all one million sites found on the web? What is the use of an IRS that mechanically indexes the information resources without bothering about the semantic value of the approach points or retrieved resources? Let us hope that the fully developed Semantic Web (Berners-Lee, 1998) would provide required solutions to the present problems.

Meanwhile, we also have subject-organised directories that endeavor to be appropriate tools for a more precise and relevant search. Isaacs (1997) opines that "In looking for the appropriate tool for a search, it may be useful to note that subject-organised directories have the following features: Context-based searching, Selected resources, improved chance of finding quality resources, and Low risk of duplication and redundancy. In contrast to automatic index searching, selective human-compiled lists will not normally throw up multiple hits for the same work." But, regarding these search tools the studies reveal that the users have negative attitude towards them. Monopoli and Nicholas (2001) state that "Direct searching is the most favourable method of collecting information. This is presumably because we are largely talking about an information-knowledgeable group of users. The ubiquitous 'keyword' search proved to be the most popular." They further state that "The primary reason given is that information seekers are unwilling to attempt new literature searching practices until they are totally convinced of their efficacy".

Whatever may be the type(s) of the search tools, in order to conceive the nature, components and form of a user-centred IRS a small study is under taken in order to find out, in its own limited way, the diversity of user approach, which is very essential to stipulate the nature of a user-centred IRS.

2 Methodology

A survey of the user approach is conducted in order to find out the diversity in user approach. A stratified sample of 25 Teachers (University Professors) and 75 Research Scholars were selected using simple random sampling technique for selection of the sample from each stratum. Out of which 4 teachers and 12 research scholars did not respond. However, the rate of response was 84% from Teachers and 84% from research scholars. A simple questionnaire containing 3 titles/topics were presented with varying degrees of complexity. The first title was fully in general terms, the second one was partially in general terms and the third one was almost full of potent key words. The respondents/users were asked to mention what keywords they use to search for information on the given topics. In addition they were asked to present the string/syntax of the keywords they have used and also what subject headings they use for information retrieval in addition to the keywords.

3 Rate of diversity/variation in the use of keywords, syntax and subject headings

The data collected through the questionnaire were tabulated and analyzed. The following Tables present the data on variation in the keywords used by the users and the rate of use of correct syntax and subject headings the users used.

Table 1. Rate of variation in the use of keywords

		Similar	%	Different	%	No	%	Similar to others	%
Topic I	Teachers	2	9.52	18	85.72	1	4.76	6	28.57
	Res.scholars.	12	19.04	51	80.95	0	0	30	47.61
	Total	14	16.66	69	82.14	1	1.19	36	42.85
Topic II	Teachers	5	23.8	13	61.9	3	14.28	N T*	
	Res.Scholars.	27	42.85	33	52.38	3	4.76	N T*	
	Total	32	38.09	46	54.76	6	7.14	N T*	
Topic III	Teachers	5	23.8	13	61.9	3	14.28	N T*	
	Res.scholars.	42	66.66	21	33.33	0	0	N T*	
	Total	47	55.95	34	40.47	3	3.57	N T*	

*Not Taken for the topics as these topics were presented with clear and potent keywords.

Table 2. Rate of use of correct syntax/string of the keywords used

		Sequence(s)					
		Right	%	Wrong	%	No	%
Topic I	Teachers	3	14.28	18	85.72	0	0
	Res Scholars	3	4.76	51	80.95	9	14.28
	Total	6	7.14	69	82.14	9	10.71
Topic II	Teachers	2	9.52	18	85.72	1	4.76
	Res Scholars	3	4.76	48	76.19	12	19.04
	Total	5	5.95	66	78.57	13	15.47
Topic III	Teachers	2	9.52	18	85.72	1	4.76
	Res Scholars	3	4.76	48	76.19	12	19.04
	Total	5	5.95	66	78.57	13	15.47

Table 3. Rate of use of Right subject headings

		Subject Headings					
		Right	%	Wrong	%	No	%
Topic I	Teachers	5	23.8	8	38.09	8	38.09
	Res Scholars	6	9.52	51	80.95	6	9.52
	Total	11	13.09	59	70.23	14	16.66
Topic II	Teachers	2	9.52	13	61.9	6	28.57
	Res Scholars	3	4.76	51	80.95	9	14.28
	Total	5	5.95	64	76.19	15	17.85
Topic III	Teachers	1	4.76	15	71.42	5	23.8
	Res Scholars	3	4.76	51	80.95	9	14.28
	Total	4	4.76	66	78.57	14	16.66

A brief analysis of the above data indicates that there is considerable diversity in user approach. The diversity varies from 33.33% to 85.72% (from Table 1). Even when we consider their approach from mutual similarity only 42.85% of them have similar approach. This shows that when the diversity is more than 50% it is very much considerable and a user-centred IR should endeavour to satisfy such a diverse approach of the users.

Coming to the syntax of the keywords, not even 20% of them either follow or know the rules of syntax. The diversity is around 85% (Table 2). When it comes to use of subject headings, hardly 15% of the users have an idea of subject headings. This shows that a user-centred system should be fully instructive, interactive and user-friendly by providing extensive and exhaustive “Help” to the users, especially when it uses controlled vocabulary or standard subject headings.

4 Relevance factor in present Machine retrieval systems

In an effort to provide everything to the users and satisfy every approach, the machine retrieval systems are violating the very purpose or philosophy of IRS. It is simply because, the machine indexing clearly proves that it lacks intellectual approach in indexing to satisfy the semantic value of user approach. The following examples illustrate how machine retrieval results in absolute rate of irrelevance and devoid of the capability of satisfying the law of 'Save the time of the user.' (Ranganathan, 1931).

A search was conducted on the Net, on December 6th 2006, using "Google" search engine to retrieve information on 'what happens when dog bites a man?' A query in natural language was used in the following fashion "What happens when dog bites a man?" The search did not match any documents. The search with *Dog bites man* retrieved 1,290,000 entries. The same search phrase with restricted search "Dog bites man" retrieved 338,000 entries/documents. The search with *Dog bite man* retrieved 1,450,000, the restricted search retrieved 139 entries. Whereas, search with Dog bite retrieved 1,610,000 and the restricted search results were 1,130,000 entries. It is so interesting to note that in IRSs that use controlled vocabulary there will not be any entries with a single letter variation such as "dog bites man" and "dog bite man", hence almost all entries will be totally relevant to the subject. Whereas the machine retrieval system shows such great difference that single letter variation results in a difference of 140,000 entries/documents. Further, the most interesting thing to note is that Google retrieved documents even on a query *Man bites dog*, the results were 1,310,000, with restricted search 356,000 entries. *Man bite dog* retrieved 1,510,000, restricted search results were 510. When it is 510 documents the user may feel that the relevance rate is very good. But none of these documents really give information about what will happen when man bites a dog?

One more search with alternative words on another subject retrieved the following results. A search was conducted on the Bilateral Relations between India and China with the following search options resulted in the following search results: Diplomatic Relations between India and China (1,110,000) ; "Diplomatic Relations between India and China" (99) ; Bilateral Relations between India and China (1,100,000) ; "Bilateral Relations between India and China" (128) ; Foreign Relations between India and China (4,520.000) ; "Foreign Relations between India and China" (2) ; Tactful Relations between India and China (54,300) ; "Tactful Relations between India and China" (did not match any documents); Relations between India and China (26,800,000) ; and "Relations between India and China" (14,500). These results show that the IRS does not have appreciable rate of Relevance. In fact, all the search elements should have retrieved same number of documents with a variation of 10% to 20% but the variation is too much we can understand the high rate of variation with open-ended keyword search. But with restricted search the variation ranges from "No matches to 128 documents with specific search elements, and no matches to 14,500 documents between specific and slightly general approach.", "510 to 356,000 documents in case of the former search." This rate of variation reveals that there is lesser sense in such retrieval and the large number of irrelevant documents retrieved consume tremendous amount of the time of the use and it is practically not possible for any user to look into all the retrieved documents.

Carlson (2003) deplores that "Information overload will go on being perceived by users as a problem, it will continue to induce feelings of stress and it will quite certainly remain a problem in terms of retrieval precision and recall." Bates (1989) likens online seeking to berry

picking. Quoting Bates, Kalbach (2000) opines that "Human information seeking behaviors in online settings present some unique problems and situations".

By and large with the above it is quite clear that user-centred IR with all virtues and precisions may be a myth. However, we need to strive to satisfy the user approaches and needs without sacrificing the performance criteria for IRSs v. gr. recall, relevance, precision, effort, response time, informativeness and form of search output, etc. as identified by Cleverdon (1966), Lancaster (1968), Perry, Kent and Berry (1956).

To be a meaningful IRS there should be a system with combined methods v. gr. machine indexing combined with manual inputs of some of the search elements especially with regard to subject indexing or provide the machine with that artificial intelligence to identify the specific subject of the document(s). An IRS with the above feature and the following might prove to be more user-centred.

5 Criteria for a User-centred IRS

On the basis of the findings of the above presented study and earlier studies conducted by Cleverdon, Lancaster and others, it may be stated that when the digital information resources in the digital era are available to distant information seeker, as there will be no scope for personal aid extended to the users, as it used to be in the conventional libraries and information systems while they were using the manual search systems, the user-centred systems of IR in the Digital Era should take the diverse approaches of the users and satisfy their approaches by providing the same search results when they search for the same subject with different keywords/synonyms/alternate terms, etc. not as it is with the present IRSs. Further, they should endeavour to be user-centred by satisfying all the natural language queries the users use. Because, most of the users may not be in a position to follow or use the rules of the syntax, controlled vocabulary, etc. When such rules or vocabulary is used the user should be provided with exhaustive guidance, help, and instructions. To put the whole thing in a nutshell, the user-centred IRSs in the Digital Era should possess the following qualities:

1. It should be a composite or combined system with both machine and manual methods.
2. It should integrate both natural language and controlled vocabulary. In addition to these, it should also be:
3. User-friendly,
4. Interactive,
5. Instructive (not merely indicative but also prescriptive and informative when controlled vocabulary is used),
6. Capable enough to meet the diverse approaches of the users by providing uniform search results,
7. Comprehensive/exhaustive (consisting of exhaustive authority files on subject headings/keywords, etc.),
8. Capable enough to do exhaustive and correct indexing,
9. Capable enough to provide poly or multiple search options v. gr. boolean, free-text, truncated, field and hierarchical searching (Perreault, 1986), and
10. Capable enough to ensure better Recall, Relevance, and Precision rates with other criteria such as effort, response time, informativeness and form of search output.

6 Conclusion

With ever-increasing capabilities of machine processing, artificial intelligence, the Information search options, capabilities can be improved by developing complex and integrated IRSs to satisfy all the approaches of the users. Then most of the IR problems can be thoroughly solved, especially, when machines understand the contents/messages the other machine(s) disseminates as envisaged by Berners-Lee (1998), there will be more intelligent and intellectual searches conducted by the machines. However, the whole philosophy is to develop user-centred and user-friendly IRSs.

References

AUSTIN, D. *PRECIS : a manual of concept analysis and subject indexing*. London : The Council of the British National Bibliography, 1974.

BATES, M. J. The design of browsing and berry picking techniques for the online search interface. *Online Review*, 1989. vol. 13, p. 407-424.

BERNERS-LEE, T. *Semantic Web roadmap*, 1998 [electronic resource]. <http://www.w3.org/DesignIssues/Semantic. html>. [Consulted: 10 dec. 2006]

CARLSON, C. N. Information overload, retrieval strategies and Internet user empowerment. In: COST 269 (Helsinki: 2003). *The good, the bad and the irrelevant: proceedings*. Vol. 1, n. 1, p.169-173.

CLEVERDON, C. W.; MILLS, J.; KEEN, E. M. *Factors determining the performance of indexing systems*. Cranfield : Aslib – Cranfield Research Project, 1966.

FERRANDANE, J. E. L. Analysis and organization of knowledge for retrieval. *Aslib Proceedings*, 1970, vol. 22, n. 12, p. 607-616.

FOSKETT, A. C. T*he subject approach to information.* London : Clive Bingley, 1982.

ISAACS, M. *TERENA Guide to Network Resource Tools*, 1997 [electronic resource]. <http://kizi2.vse.cz/~jjkastl/GNRT/websearch/directories.html>. [Consulted: 12 dec. 2006].

KALBACH, J. Designing for information foragers: a behavioral model for information seeking on the World Wide [electronic resource]. *Web Internetworking*, 2000, vol. 3, n. 3. http://www.internettg.org/newsletter/dec00/article_information_foragers.html. [Consulted: 17 aug. 2006].

KENT, A. *Information analysis and retrieval*. New York : Becker and Hayes, 1971.

LANCASTER, F. W. *Information retrieval systems: characteristics, testing, and evaluation.* New York: John Wiley & Sons, 1968.

MONOPOLI, M.; NICHOLAS, D. A user-centred approach to the evaluation of subject based information gateways : case study SOSIG. *Aslib Proceedings*, 2000, vol. 52, n. 6, p. 218-231.

PERREAULT, J. M. Some perils of the “user-friendly” attitude in cataloguing. *Advances in Librarianshi*, 1986, vol. 14.

PERRY, J. W.; KENT, A.; BERRY, M. M. *Machine literature searching*. New York: Interscience, 1956.

RANGANATHAN, S. R. *The five laws of library science*. 2nd ed. reprinted. Bangalore: Sarada Ranganathan Endowment for Library Science, 1988. (First edition was published in 1931).

SHARP, J. R. *Some fundamentals of information retrieval*. London : Deutsh, 1965.

TAUBE, M. et al. The uniterm coordinate indexing of reports. *The Technical Report.* New York: Reinhold, 1954.

Acknowledgements

I thank Mr. K. Purnachandra Rao, Research Scholar in the Dept. of Library and Information Science, Osmania University for his help in preparing this paper. I also thank all the Teachers (Professors) and Research Scholars of University College of Arts and Social Sciences and other colleges of Osmania University for having responded to the questionnaire and provided me with necessary data.