International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 4, Number 3 (2014), pp. 221-230
© International Research Publications House
http://www. irphouse.com /ijict.htm

Refinement of Web Search using Word Sense Disambiguation and Intent Mining

Pooja Bassin[1] and Manisha[2]

[1] M.Tech Scholar, Department of Computer Science, Banasthali University,
[2]Department of Computer Science, Banasthali University, Jaipur, India.

Abstract

Searching or retrieving information on the Internet using any web search engine displays documents in enormous amounts. Information Retrieval is the activity of tracing and obtaining relevant information from a collection of information resources. It is a challenging task for retrieving relevant information over the Internet. Although there are search engines to help yet they display information in astronomic amounts. Search engines display results in form of documents. The output results do not give us guarantee of relevant information. Therefore it becomes the user's task to dig out relevant information out of the bulk of results. This is the major research issue in Information Retrieval. This problem can be reduced by the use of intent mining. Hence to determine the query intent is an important problem for today's search engines. The major task is to comprehend the user's intention behind the query he has entered to the search engine. Users may be searching for an explicit page, browsing for information, or just trying to purchase goods. Guessing the impeccable intent is important for returning apt results. Queries can be as short as one word and consist ambiguous terms. Search engines need to derive what users want from this narrow source of information. Understanding the intent behind a user's query can help search engines to exhibit results that are suitable hence saving user's time from filtering out relevant documents out of irrelevant ones, thus, greatly improving user satisfaction.

Keywords: Word sense disambiguation; intent mining; information retrieval; web search.

1. Introduction

An extensive range of information is available on the Internet. This information is not defined in any structured form as well as the Internet does not have centralized governance for access or usage of information. Hence the user searching information on the Internet will come across such results which won't be of any use to him. The user himself will have to either filter out results which are relevant to him or he will have to reformulate his query unless he reaches a set of results which he might find apt. The queries entered by the user can be as long as 32 words which is an applicable limit by Google or as short as one word. Such queries can have words which can be ambiguous in nature i.e. a word in English language that can have more than one implication for different frames of reference. Hence such short queries do not provide enough contexts to differentiate them from conflicting meanings of the given words. In our research we present such a system which helps us in refining the user query based on his intentions behind entering the query.

2. Classification of Queries

Web queries have been broadly classified into following three categories (Manning et al, 2007):

- **Informational Queries**: Queries that cover a broad topic for which there may be thousands of relevant results.
- **Navigational Queries**: Queries that seek a single website or web page of a single entity.
- **Transactional Queries**: Queries that reflect the intent of the user to perform a particular action, like purchasing a car or downloading a screen saver.

In our study we have added a fourth category of query titled as Directional Queries which can be defined as such queries that try to determine direction, path or route of a certain destination.

3. Experimental Setup

In this study we have worked on the examples of above mentioned queries including Directional queries. For word sense disambiguation, some ambiguous words such as orange, bank, crane, bass etc. have been taken into account. The system has been developed on Java platform and thus experiments have been performed based on the profile of the user.

4. Methodology

In the above figure, the user enters a query into the system. This query goes through the stemming phase that returns various stems thus they further go for: Sense Identification and Intent Identification. Both of the approaches lead to the generation of a refined query which goes to the web search engine and hence results are displayed back to the user.

The System Flow Architecture has been shown below:

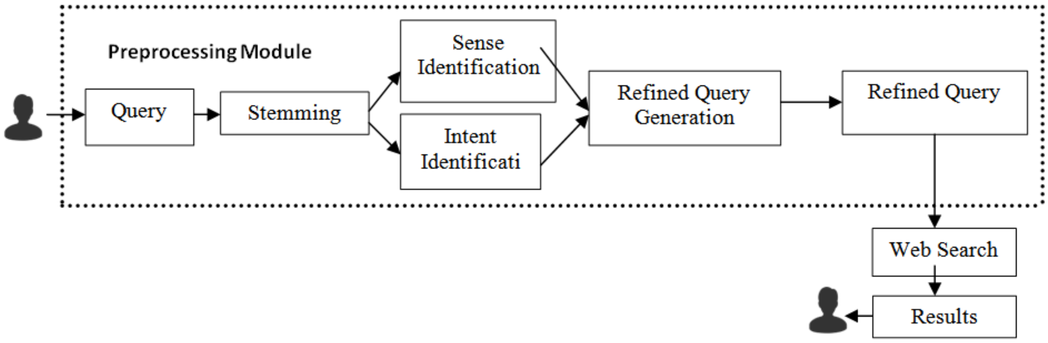


Fig. 1: System Flow Architecture.

5. Evaluation and Results

The main purpose of a program evaluation is to determine the quality of a program by formulating a judgment (Hurteau et al, 2009). In information retrieval, precision and recall are considered as standard metrics for evaluation. Our evaluation criterion is based on precision alone. Precision was calculated for four cases for each type of query and for a set of ambiguous words. The output for both original user query and the refined query after the Refined Query Generation process have been compared and thus graphs have been plotted.

Table 1: Precision table comparing two page outputs of user queries and refined queries.

Type of Queries	**Cases(C)**	**User Query Results**		**Our System Results**	
		Page 1	**Page 2**	**Page 1**	**Page 2**
Informational Queries	C1	0.07	0.00	0.83	0.90
	C2	0.53	0.36	0.78	0.80
	C3	0.50	0.38	0.92	0.83
	C4	0.35	0.45	0.71	0.72
Navigational Queries	C1	0.50	0.10	0.63	0.40
	C2	0.32	0.11	0.81	0.50
	C3	0.24	0.00	0.55	0.10
	C4	0.44	0.16	0.64	0.20
Transactional Queries	C1	0.26	0.29	0.68	0.70
	C2	0.33	0.44	0.56	0.71
	C3	0.55	0.36	0.63	0.53
	C4	0.63	0.42	0.77	1.00
Directional Queries	C1	0.08	0.00	0.60	0.50
	C2	0.14	0.00	0.40	0.66
	C3	0.14	0.10	0.33	0.10

	C4	0.13	0.09	0.66	0.30
Sense Identification Set	C1	1.00	1.00	0.20	0.75
	C2	0.53	0.38	1.00	0.75
	C3	0.11	0.09	1.00	0.90
	C4	0.45	0.20	1.00	1.00

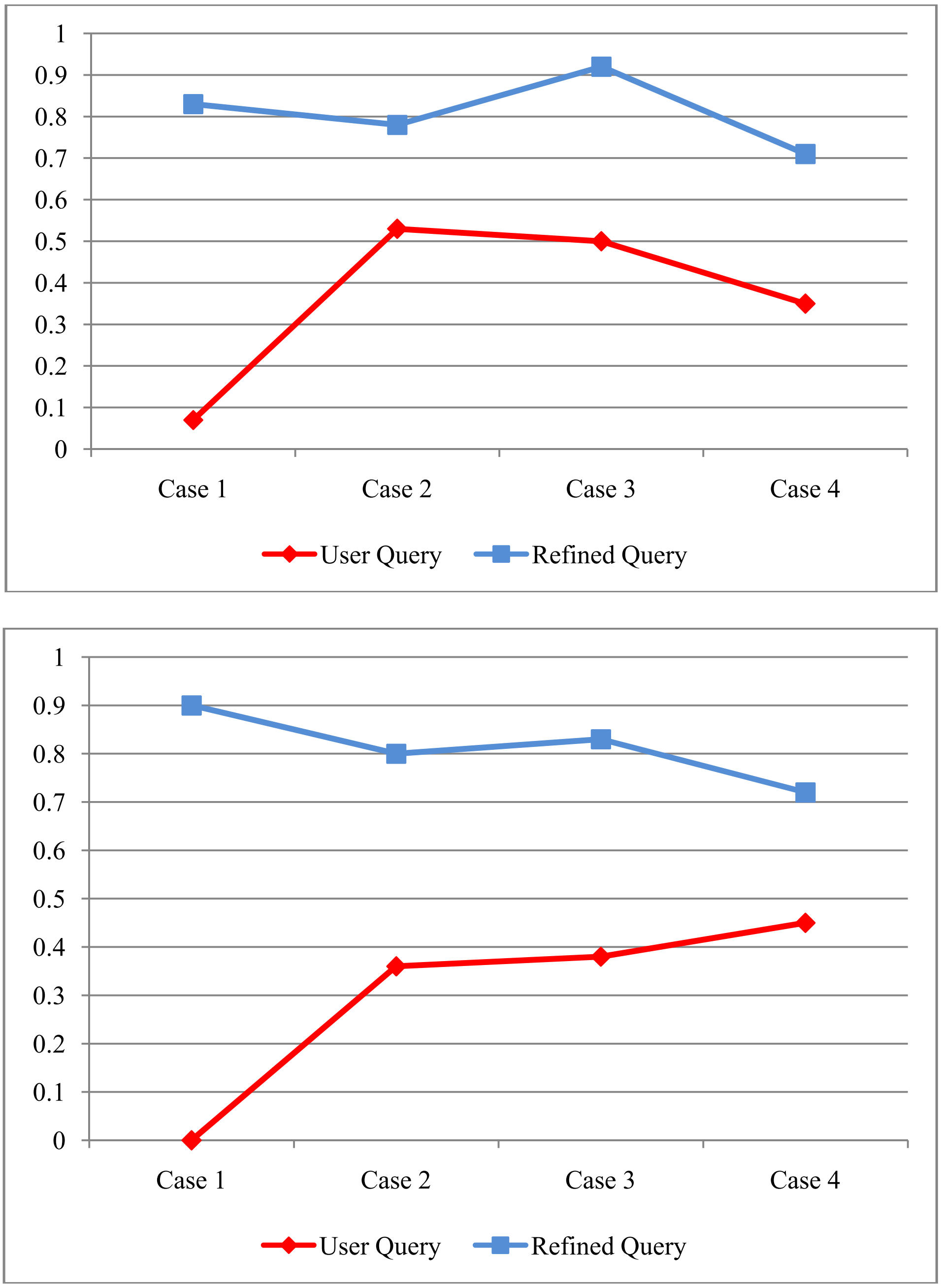


Fig. 2: Precision graphs for Informational Queries for page 1 and page 2.

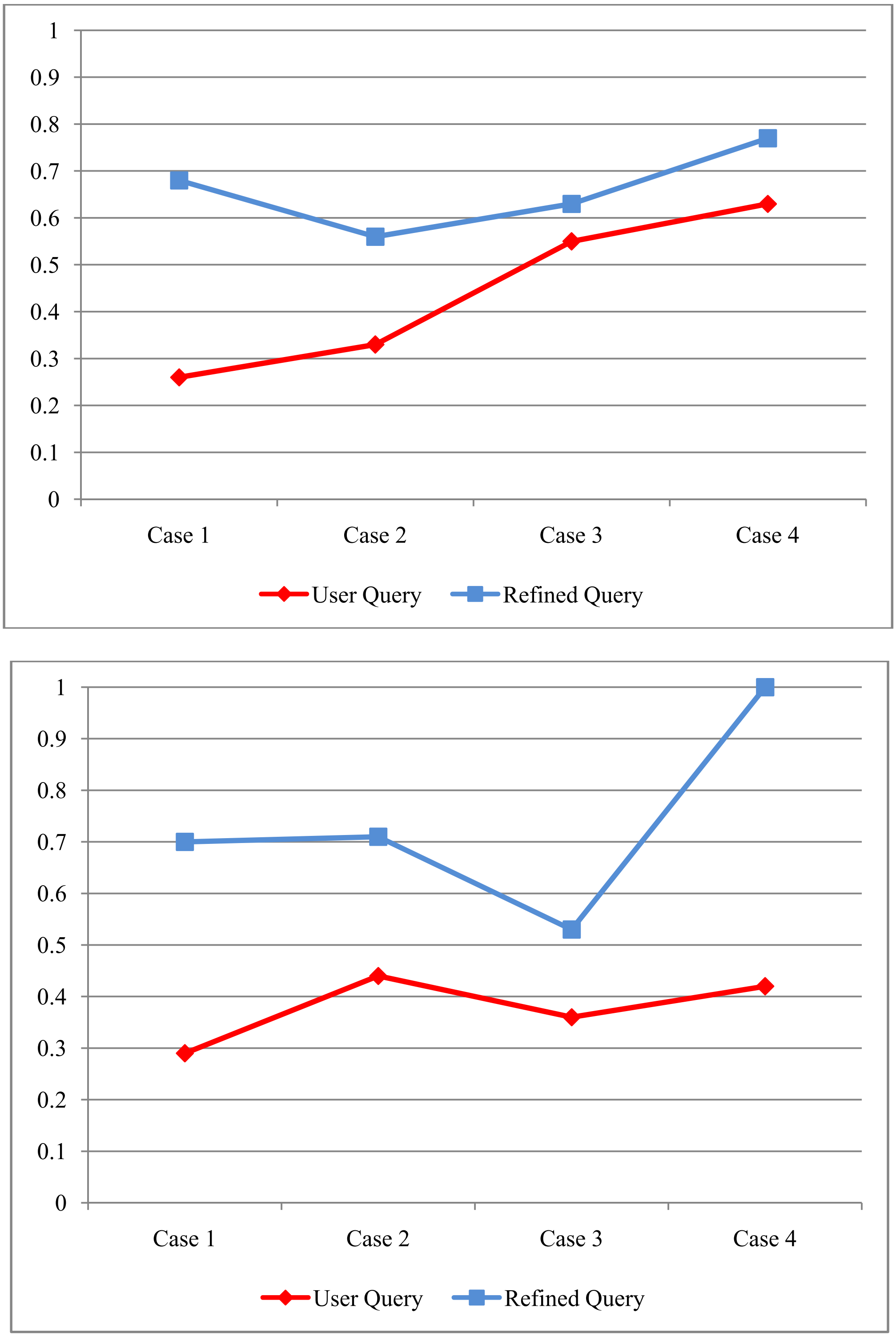


Fig. 3: Precision graphs for Transactional Queries for page 1 and page 2.

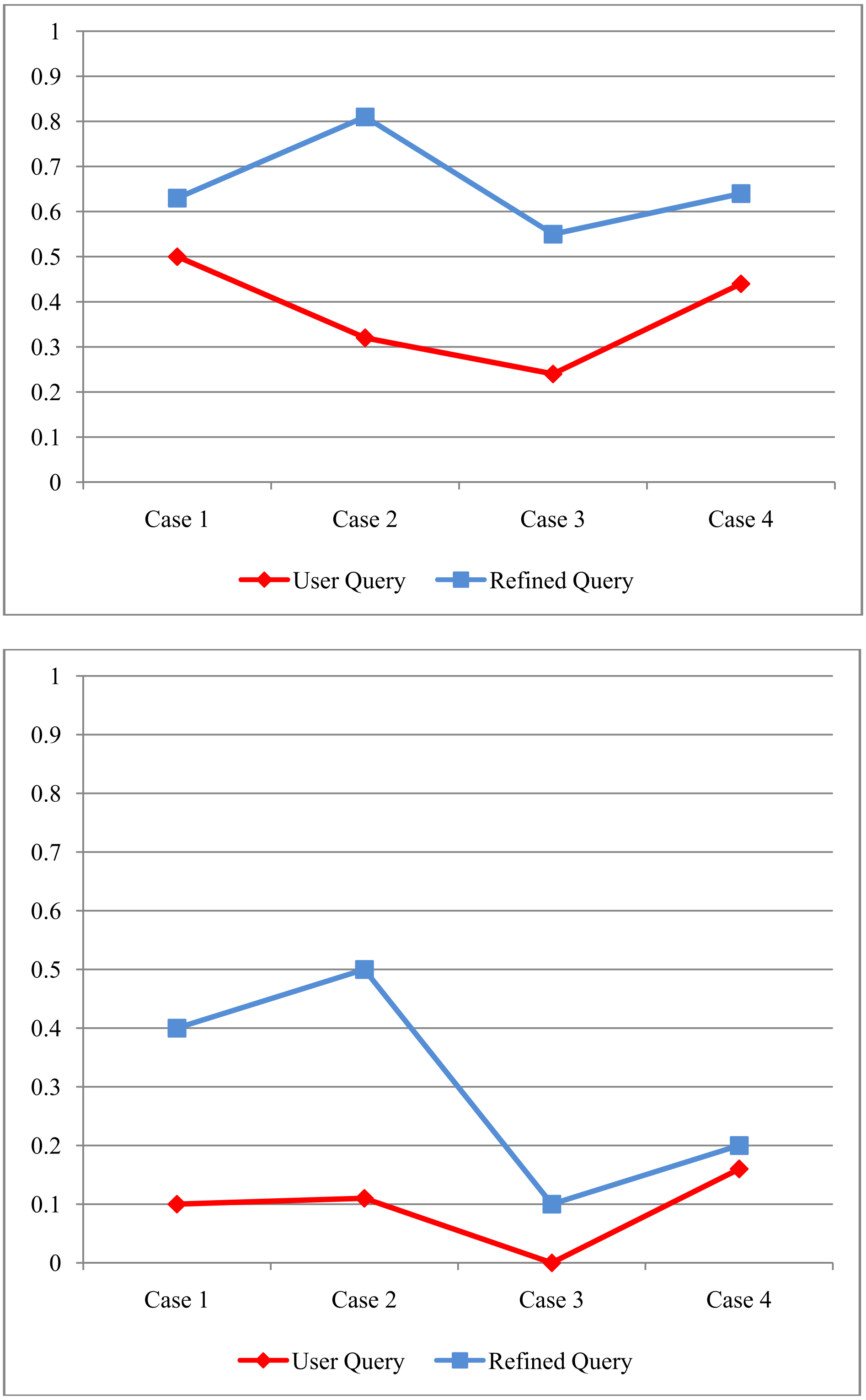
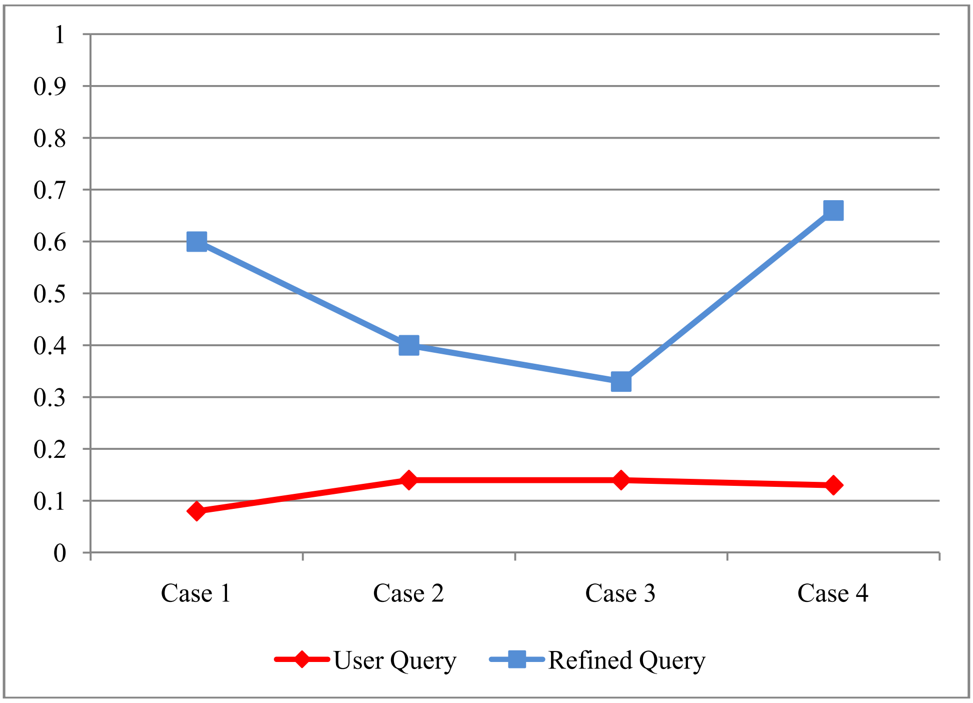


Fig. 4: Precision graphs for Navigational Queries for page 1 and page 2.



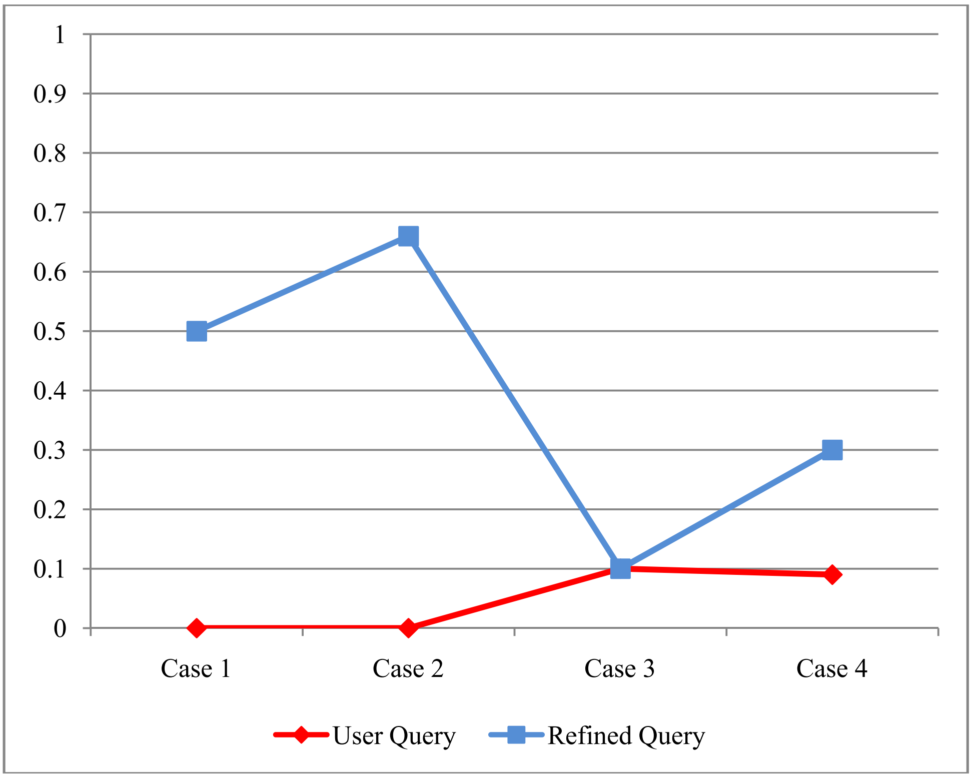


Fig. 5: Precision graphs for Directional Queries for page 1 and page 2.

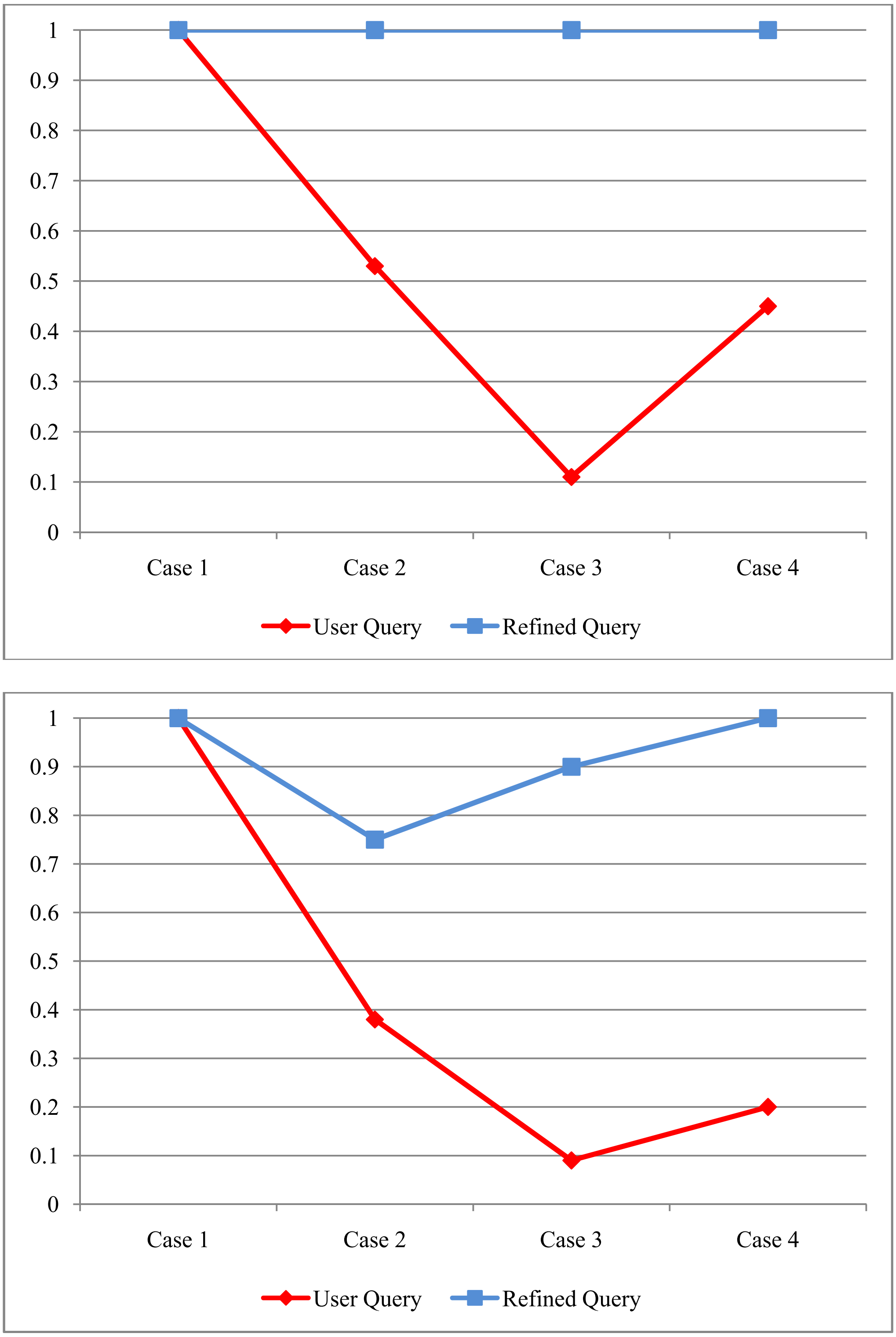


Fig. 6: Precision graphs for Sense Identification for page 1 and page 2.

Corresponding precision graphs for the queries shown in the table have been shown above. For the given precision table and precision graphs the formula which has been used for precision is:

Precision = | {relevant documents} | ∩ | {retrieved documents} |
| {retrieved documents} |

Hence all the results have been evaluated based on the above formula.

6. Conclusion

In this paper we have described a methodology which would help us not only to display results based on keywords entered by the user but also based on his interests and background for whatever he will be looking for on the Internet will be exhibited to him. By seeing the precision table and graphs it can be concluded that such a system would bring out better results rather than the simple search based completely on keywords entered by the user. The intention behind typing the query should be scrutinized and thus better outcomes would be generated.

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze (2007), Ch. 19
http://www.wordstream.com/blog/ws/2012/12/10/three-types-of-search-queries

[2] Hurteau, M.; Houle, S., & Mongiat, S. (2009). "How Legitimate and Justified are Judgments in Program Evaluation?" *Evaluation* **15** (3): 307–319

[3] Powers, David M W (2007/2011). "Evaluation: From Precision, Recall and F-