Factual or Satisfactory: What Search Results Are Better?[1]

Yu Hong, Jun Lu, Jianmin Yao, Guodong Zhou, Qiaoming Zhu
[a]School of Computer Science and Technology, Soochow University,
No.1 Shizi Street, Suzhou City and 215006, China
{hongy, 20104227066, jyao, gdzhou, qmzhu}@suda.edu.cn

Abstract. An interesting issue “whether the factual search result is better than the satisfactory one?” is discussed in this paper, and all effort is to illustrate the publicly recognized conception “the supreme satisfactory result is optimal” could lead Information Retrieval astray. By contrast, we propose a new hypothesis “the factually result should be the optimal although sometimes unsatisfactory”. To implement the pilot study on this, we developed a search-engine (Google) based labeling platform and designed a new evaluation method involving user experiences. The results on the platform show the existence of bias on determining the optimal search results (factual or satisfactory). And the very different performances of NDCG and our evaluation metric demonstrate the weakness of current ranking algorithms in mining and recommending the long-term effective information.

Keywords: Factual search result, satisfactory search result, information retrieval, evaluation

1 Introduction

What services are current search engine providing? The universal search engines are mining and recommending all possible relevant information to users. Further, the personal search engines are learning the real intention of users and recommending the information that extremely satisfies the personal requirements. But whether are the satisfactory search results factually factual? The answer should be no. In detail, it can be sure that users aim to explore the unknown when using the search engine, and thus they should have little or even no prior knowledge to support their judgments on the satisfaction with the search results. This may result in the probability that any result with a little relation to the query (even wrong one) can be determined to be satisfactory. It just like that a lost visitor would always say “thank you” to the helpers even if he was shown the wrong way.

Both the explicit and the implicit feedbacks can reflect the user experience of search results (satisfied or not). So the feedbacks are normally used to adaptively learn user intention and iteratively improve the retrieval performance. However, as the discussion above, the judgment on the satisfaction might be imprecise. Thus the adaptive learning, as the most important component of personal search engines, is possibly using the unreasonable feedbacks to mislead the retrieval process.

The doubts mentions above raise two questions: 1) Whether are the satisfactory results always the helpful information for users to acquire knowledge? And whether can they consistently enhance the intelligence of information retrieval? 2) If the answer is no, whether does there exist other kind of information to compensate for the potential failings of users even or supersede the satisfactoriness?

In this paper, we propose a hypothesis that the better search results should be the factual information. The factualness mentioned here is not “users think something right or find it nice” but means the nature of information is really right. The best factual information should be the theorem, scientific laws, and all the knowledge that meets with the law of nature, e.g. “the

[1] Supported by National Natural Science Foundation of China (Grant No. 60970057, 61003152) and Natural Science Foundation of Jiangsu Province Suzhou City (Grant No. SYG201030)

reason why an apple fell on Newton's head is because of the universal gravitation". Besides, the factual information should also include the centencyclopedism, common-sense knowledge and any principle that has been proved for centuries by human, e.g. "Not eating roughage will cause beriberi". The hypothesis supports the view that if we know the users' intention, we shouldn't indulge in satisfying their individual preferences or interests, but supply the tried and true knowledge for them to capture the essence of things.

The hypothesis implies a new way of knowledge acquisition: factualness-based information retrieval. And it should present new challenges to current retrieval techniques:

- Identify and mine the actual knowledge in large-scale information resource.
- Detect the target of knowledge acquisition and establish the corresponding query model.
- Involve factualness measurement into retrieval model.

But in this paper, as a pilot study on this issue, we focus on proving the hypothesis, viz., the factual search results should be better than the satisfactory ones. For this, we developed a search-engine (Google) based labeling platform and designed a new evaluation method to measure the precision along with user experiences. The results show strong differences of user opinions on satisfactoriness and factualness, and the factualness has more direct influences on user mood.

The paper is organized as follows: Section 2 discusses the difference between factual and satisfactory search results. Section 3 introduces the labeling platform. Section 4 gives the main labeling results. Section 5 gives the factualness-based evaluation metric and shows the performance of current retrieval model on acquiring the factual information.

2 Difference Between Factualness and Satisfactoriness

The factualness should be different from the relevance or satisfactoriness. For instance, suppose a query is "what is the Chinese Dragon", how can we classify the following search results (relevant, satisfactory and/or factual):

- *1) Do you want to know Chinese Dragon, see this book "Chinese History and Culture"*
- *2) Recently some researchers had proved the Chinese Dragon is but the crocodile*
- *3) Chinese Dragon is the totem of the Chinese nation*

Obviously, 1) is the relevant but not satisfactory information because it cannot reduce the effort of users in getting the answer. On the contrary, 2) normally can satisfy users by the forthright and seemingly authoritative answer although it is wrong, 3) is satisfactory if users have enough knowledge of the Chinese culture, otherwise and actually often it is undesirable although factual.

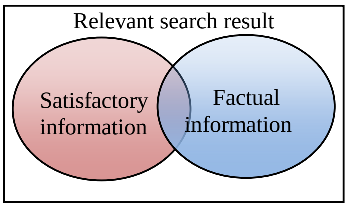

Figure 1: The intersection between factual and satisfactory search results.

It can be believed that both factual and satisfactory information normally are relevant search results. Therefore it is unreasonable to affirm that a search result is consequentially either factual or satisfactory. Actually they have the intersection as shown in the Figure 1. Obviously the search results in the intersection are the optimal because the satisfactoriness can persuade users to accept the knowledge in the results and meanwhile the factualness can ensure the knowledge is right. Thus the difference should occur in the two parts outside the intersection:

- Factual but not satisfactory (abbr., FNS) search results.
- Satisfactory but not factual (abbr., SNF) search results.

The main characteristic of FNS is that the results conflict with the preference, interest, sentiment or prior cognition of users. For example, if the query is “what is my test scores?”, the following factual answers will receive different satisfactoriness (suppose the full mark is 5):

- *Your score is 5 (Satisfactory).*
- *Your score is 1 (Dissatisfactory).*

On the contrary, the main characteristic of SNF is that the results conflict with the truth. For example, if the query is “Daddy, will Santa Claus come tonight?”, the satisfactory answers will involve different factualness:

- *Yes, but don’t be surprised if he looks like me (factual).*
- *Yes, he will drive sled to our roof (untrue).*

As the discussion above, the search results in the intersection are obviously optimal. But which one of FNS and SNF is better? In the rest sections, we focus on introducing a google based labeling platform and the experiment results on it, by which to prove the initial hypothesis: factual information can improve user experience much more.

3 Factualness Labeling System and Results

In this section, we firstly introduce our search-engine based labeling platform for information factualness and satisfactoriness, and then we give the labeling results along with the corresponding analysis.

3.1 The Labeling Platform

The labeling platform includes three components: search engine, interaction interface and database. The search engine provides the service of information retrieval by remotely accessing Google. The interaction interface is used to display the search results and acquire user experiences. And the database is used to preserve the prepared queries and the corresponding factual results, and besides, it is also used to record the user experiences in real-time. Here we focus on introducing two key issues:

- How to obtain the queries having the corresponding factual results?
- How to collect the user experiences?

Considering the necessity for the first issue to ensure the factualness of the results, we regard encyclopedic knowledge of the Chinese language (We have only Chinese volunteers) as the best resources to obtain the queries. We extracted 279,576 items of knowledge from the Wikipedia and used the titles as the queries. Correspondingly the web pages which involve the knowledge are used to be the factual results. After filtering the ordinary and long queries which caused the difficulty for the factual results to be ranked at the top of the result list (users seldom can see them), we got 3,719 queries.

To ensure the realistic user experiments, we display the prepared queries on the interface for users to select the favorite ones. Also we make the displayed queries change in real-time to increase the likelihood of users finding their interests. To obtain the user experiments, we designed two forms to keep records of users’ opinions on factualness, satisfactoriness and their mood for each clicked feedback. One form (FBC) includes 6 questions and opens before the click (see Table 1), the other form (FAC) includes 8 questions and opens after the click (see Table 2). The FBC is used to record the initial user experiences on a specific feedback, on the contrary, the FAC record the experiences when the users have gone deep into the details of the feedback. On the basis, a complete labeling process should be: Step 1: select and click a favorite query on the interface; Step 2: browse the feedbacks and select one of them; Step 3: fill out the form FBC and click the feedback; Step 4: fill out the form FAC; Step 5: go back to Step 2 or exist.

Table 1: The questions that should be answered before clicking a feedback.

Form before Click (FBC)
Q1: Whether is it the first feedback? *(Y/N)*
Q2: Whether have you the habit to neglect the first feedback and/or directly check the feedback on the current position? *(Y/N)*
Q3: Have you checked the feedbacks before this? *(Y/N)*
Q4: How satisfied are you with this feedback? *(Radio button)*
A. Very (5)   B. just satisfied (4)   C. A little (3)   D. Hard to say (2)   E. Dislike it (1)
Q5: Why did you check this feedback? *(Radio button)*
A. Must be factual (5)   B. Should be factual (4)   C. The abstraction looks right (3)   D. Rank is high (2)   E. At random (1)
Q6: How are you feeling now? *(Radio button)*
A. Feel good (5)   B. Be interested (4)   C. Struggle to continue (3)   D. Anxiously (2)   E. Be wearisome (1)

Table 2: The questions that should be answered after clicking a feedback.

Form after Click (FAC)
Q7: How satisfied are you with this feedback? *(Radio button)*
A. Very (5)   B. just satisfied (4)   C. A little (3)   D. Hard to say (2)   E. Dislike it (1)
Q8: How sure are you about the factualness of the feedback? *(Radio button)*
A. Very sure (5)   B. Sure (4)   C. Maybe (3)   D. Hard to say (2)   E. It is wrong (1)
Q9: How sure are you about the wrongness of the feedback? *(Radio button)*
A. Very sure (5)   B. Sure (4)   C. Maybe (3)   D. Hard to say (2)   E. It is factual (1)
Q10: What makes you think it is factual? *(Radio button)*
A. High authority   B. Credible source   C. Credible contents   D. Seems reasonable   E. I have the same thought
Q11: What satisfies you? *(Radio button)*
A. Provide everything I need (5)   B. I can learn some knowledge (4)   C. Interesting (3) D. Beautiful webpage (2) E. others (1)
Q12: Please rank all the results you have seen in the order of factualness.
Q13: Please rank all the results you have seen in the order of satisfactoriness.
Q14: How are you feeling now? *(Radio button)*
A. Feel good (5)   B. Be interested (4)   C. Struggle to continue (3)   D. Anxiously (2)   E. Be wearisome (1)

4 Main Labeling Results

There are 19 volunteers trained to use the labeling platform, and they submitted 1,659 FBC-click-FAC forms for 719 queries in a month. After filtering the incomplete forms and the errors caused by faulty operation, we finally got 1,238 forms for 539 queries.

Before the discussion on the main labeling results, we firstly give the frequently-used terminologies:

- FBC-click-FAC form: correspond to a click and include both FBC and FAC
- FBC form: a form need to be filled out before users clicking a feedback
- FAC form: a form need to be filled out after users having gone deep into the details of a feedback
- Patience level: an indicator of patience with the labeling process for a query. Here, we use the number of the FBC-click-FAC forms completed by users to indicate the level. For example, given a query, if users click 2 feedbacks and successfully submit the corresponding 2 FBC-click-FAC forms, the Patience level will equal to 2.

Additionally, to quantify the user opinions on the questions of factualness, satisfactoriness and experiments in the FBC-click-FAC form, we gave the answers scores from 1 to 5 according to the intensity (see Table 1 and Table 2). Finally, the quantitative data or/and trend curves that reported below are generated for different QA combinations, so we add the notations of the combinations (e.g. Q7+Q8) into the sub-headings for better discrimination.

4.1 Basic data distribution (Q1+Q2+Q3)

According to the data of the FBC-click-FAC forms, we find the volunteers reached the factual feedbacks of only 20.52 percent of the queries. But according to the prior records in the database, there are about 50.09 percent of the queries (viz. the queries selected by the volunteers) have factual feedbacks in the top 10 of the result list, and even all volunteers reported they at least browsed the whole feedbacks in the first page (viz. the top 10). So there are good reasons to believe that the volunteers missed the factual feedbacks of 29.57 percent of the queries.

Table 3: Recalls on different patience levels.

Patience level	*Recall (%)*	*Clicks (num)*
1	19.37	191
2	19.54	266
3	25.17	429
4	31.03	116
5	28.00	125
6	0.00	90

Besides, by inspecting the recalls on different patience levels (see Table 3), we find two interesting phenomenon:

- Users should have little patience to check more feedbacks (see the column of “Clicks” which gives the number of clicks on different patience levels)
- As the patience level increases, the recall of factual feedback rises (see Figure 2)

The phenomenon illustrate that users can gradually raise the capability of recognizing the factual feedbacks if they have the patience to check and learn more information. However, the truth is that most users only check 1 to 3 feedbacks (see Table 3), thus the probability of reaching factual feedbacks and successfully recognizing them are both low, especially when the factual feedbacks are far from the top.

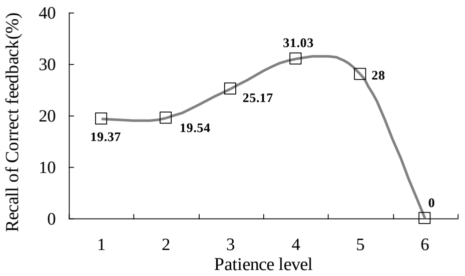

Figure 2: Recall rises as patience increases.

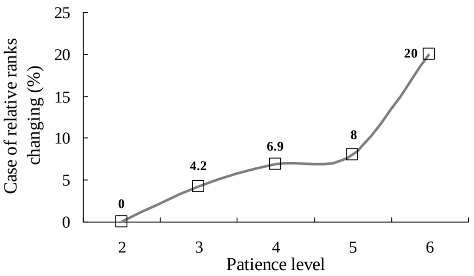

Figure 3: The probability of relative ranks changing.

4.2 Data of cognition transformation (Q12)

The basic data distribution also shows an exception that the recall weakens when the patience level increases up to 5. Then if the exception is not caused by the sparse data on the levels, a question is whether user cognition can be disturbed by over-learning. To answer the question, we inspected the “order of factualness” on different levels by using the data of Q12 in the FBC-click-FAC form. And we focus on surveying two issues:

- Whether does relative rank change as patience level increases?
- Whether will the ranks drop when users learn new things?

For the first issue, on every patience level, we calculated the average probability of relative rank changing as a new click occurs (see Figure 3). Here, the “relative rank changing” means the linear order of prior feedbacks is changed by volunteers according to their opinions on factualness. For instance, if the original order of the i-th and j-th clicked feedbacks is {i, j}, when user k-th clicks a feedback and arrange the feedbacks, within the possible orders {i, k, j}, {i, j, k} and {k, i, j}, the relative rank of {i, j} never changes, but the orders {j, k, i}, {j, i, k} and {k, j, i} change the rank.

This test is used to inspect the disturbance degree of new knowledge to user cognition on factualness. As shown in Figure 3, the disturbance does exist and its degree increases with the patience level. This phenomenon illustrates that users might be puzzled when they are confronted with lots of possible factual information. Thus, considering the little patience of users and the disturbance of over-learning, it is highly necessary for search engines to top-rank the factual information, otherwise users either miss the factual information or lose themselves in the redundant information. However, as reported in *1),* actually only 50.09 percents of factual information appear in the top 10. This is really a big problem.

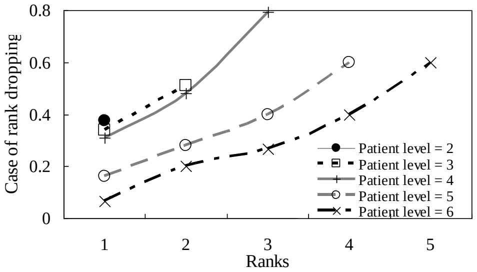

Figure 4: The probability of ranks dropping.

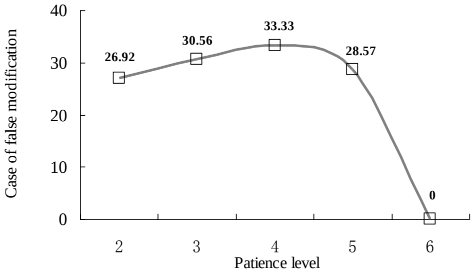

Figure 5: The probability of false modification.

For the second issue, we calculated the probability of rank dropping for every click at different patience levels (see Figure 4). Here, the "rank dropping" means the ranks of previously-clicked feedbacks is dropped by volunteers after a new click occurs. For instance, if the original order of the i-th and j-th clicked feedbacks is {i, j}, and the k-th clicked feedback is inserted into the linear order as {i, k, j}, then the rank of j-th click drops but the i-th doesn't.

This test is used to inspect the obstinacy of users to believe their opinions on factualness. As shown in Figure 4, whatever the patience level is, the probability of rank dropping at the first click is lowest. The labeling results illustrate that the users always persistently believe the feedback they initially clicked is factual. However, the factual precision of the first clicks is only 15.4%. Additionally if users cannot successfully reach the factual feedback at the first click, it will be harder for them to precisely recognize it in subsequent browsing process (see the whole precisions in Table 4). This supports the imagination discussed in the "eye tricks": users always easily accept what they learn or be taught initially and keep the cognition for a long time. But obviously current search engines seem to never consider the characteristics of users cognizing the unknown and the possible negative influence of the initial mistakes on the whole learning process.

Table 4: Precision at different clicks.

i-th click	*1st*	*2nd*	*3rd*	*4-th*	*5-th*
Precision (%)	15.40	6.32	5.12	1.39	0

Table 5: False modification at different patiences.

Patience level	*2*	*3*	*4*	*5*	*6*
False Modification	26.92	30.56	33.33	28.57	0.00

4.3 Data of false modification (Q12)

By the analysis in 1), we have found users can gradually raise the capability of obtaining the factual information if they have the patience to check and learn more search results (higher recall on high patience levels as shown in Figure 2). But it is still unknown for us how many mistakes they make before the factual click. For this, we checked the probabilities of false modification on each patience level (see Table 5). Here, the "false modification" means users mistakenly drop the rank of the factual feedback. Actually, besides of the false modification, the mistakes should also include the wrong clicks before users successfully reaching the factual feedback. But we regard the false modification as the more serious mistake because it happened after users having known the factual knowledge. It can be used to measure the confidence of users on factual feedbacks.

As shown in Table 5, nearly on all patience levels there exists false modification, and the corresponding probability increases as the patience level rise until up to 4 (see Figure 5). This illustrates the initial determination on factualness of users may often depend on intuition or indirectly relevant knowledge, which make it hard for users to confirm the factualness of information. And it can be believed that the false modification should be the main part of which causes the very low precisions on the subsequent clicks (see Table 4). But interestingly, the probability of false modification at the 2nd level is relatively low. It is actually because users always put more trust in their first selection and seldom to modify it, as the discussion in 2). Obviously it should be another challenge task for factualness-based information retrieval to identify and weaken the hesitation of users on the factual information.

4.4 Cognition consistency (Q4+Q5)&(Q7+Q8)

By the analysis in 1), 2) and 3), we concluded the characteristics of user recognizing the factual feedback: 1) Low precision and recall; 2) easy to be puzzled by wrong feedbacks; 3) persistently embrace the feedback they initially clicking; 4) be uncertain about the factualness of every determination. Thus, a necessary question is what makes the users to accept and embrace the wrong information and why they can be puzzled. As the discussion in section 1, one possibility to cause this may be the difference between the user cognitions on the factualness and satisfactoriness. In details, users may always concern with whether information meat their requirements but neglect the factualness, or/and even they cannot distinguish factualness from satisfactoriness. To verify this, we checked the answers of Q7, Q8, Q12 and Q13 in the FBC-click-FAC forms to survey the possible existence of the difference.

Table 6: Scores of satisfactoriness at different clicks.

i-th click	*SBC*	*SAC*	*Differenc*
1st click	3.183	3.646	0.517
2nd click	3.357	3.214	0.143
3rd click	4.000	2.556	1.434

Table 7: Scores of factualness at different clicks.

i-th click	*SBC*	*SAC*	*Difference*
1st click	3.366	3.556	0.068
2nd click	3.286	3.203	0.083
3rd click	2.333	2.258	0.075

We firstly inspected the average scores of satisfactoriness at the 1st, 2nd and 3rd clicked feedbacks before the clicks (Score-before-Click, abbr., SBC) and that after the clicks (Score-after-Click, abbr., SAC), see Table 6. The results show the average SBC of satisfactoriness continuously increases as the click number increases, but the average SAC continuously reduce. This illustrates the satisfactoriness is very different before and after the click. According to the difference, we can at least believe that users are always aware of what they need clearly. It is because they can completely demolish their prior opinions after go deep into the details of information.

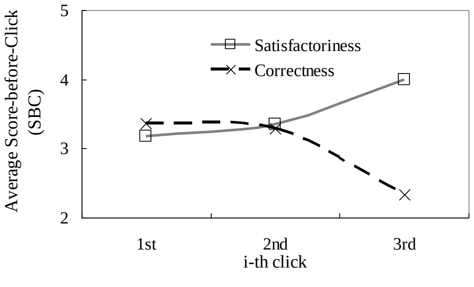

Figure 6: SBC distribution

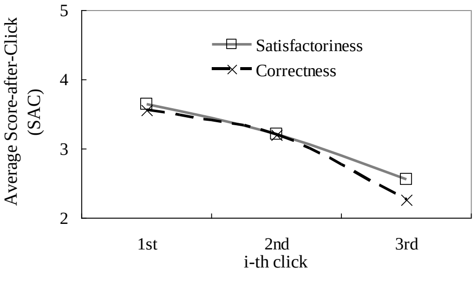

Figure 7: SAC distribution

On the contrary, the condition of factualness is very different. We inspected the average scores of factualness at the 1st, 2nd and 3rd clicked feedbacks before the clicks (SBC) and that after the clicks (SAC), see Table 7. The results show the average SBC of satisfactoriness and the SAC both continuously reduce as the click number increases. And especially, the differences between the SBC and SAC are very small. Thus, if it is right that users cannot determine the factualness by only checking the title or/and snapshot of feedback (actually the low precision has proved this), the small difference between the SBC and SAC may illustrate users normally are not aware of what is the factual information.

Secondly, we inspected the differences between the SBC of satisfactoriness and factualness at the 1st, 2nd and 3rd clicked feedbacks. The results show very different trends (see Figure 6). The rising trend of the satisfactoriness reflects the possibility that users can be easily satisfied so long as the information looks interesting, relevant or/and suiting their tastes. On the contrary, the reducing trend of the factualness illustrates that users may be more fastidious about the factual information. Thus it seems users have completely different cognitions on the satisfactoriness and factualness. However, by analyzing the differences between their SAC, we surprisingly found the cognitions are very similar (see Figure 7): both the SAC trends reduce as click number increases. This reflects the consistent meticulousness of users in evaluating the qualities of information when they go deep into the details.

As a conclusion, we can find following characteristics of users evaluating satisfactoriness and factualness: 1) Click a feedback with the low standard of satisfactoriness but high when checking the details uzzled by wrong feedbacks; 2)With high standard of factualness to select and evaluate the feedbacks Persistently embrace the feedback they initially clicking. Therefore, it can be believed that the relative random clicking creates the opportunity for wrong feedbacks to puzzle users. And if the users are puzzled at the beginning, they will easily embrace the mistake for a long time.

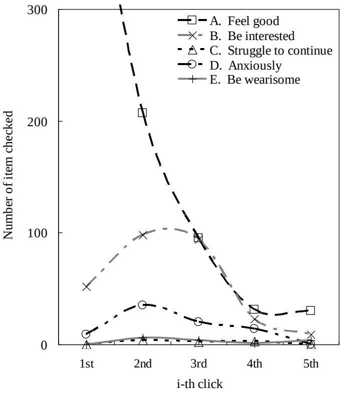

Figure 8: Mood swing before click (MBC).

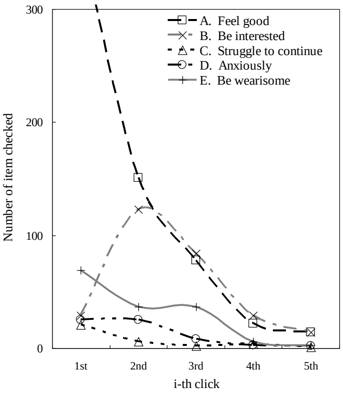

Figure 9: Mood swing after click (MAC).

4.5 Dada of mood swing (Q6+Q14)

In our experiments, we accordingly found the volunteers always have mood swings. In our imagination, the mood swings should be caused by the satisfactoriness or/and factualness. To verify this, we inspected the answers to Q6 (Mood-before-Click, abbr., MBC) and Q14 (Mood-after-Click, abbr., MAC) and counted the number of marks at from the 1st to 5th clicks. The whole trends of mood swings on the clicks are shown in Figure 8 (MBC) and Figure 9 (MAC). Here, it is necessary to note that the reducing trend of each curve in the figures should be neglected because the practical number of clicks at the higher patience level is really small. Thus the factual perspectives should be the height of the whole curve in the two-dimensional space and the height difference among the curves. By comparing the two figures, we found two characteristics: 1) The differences of curves A, B, C and D (viz., feel good, be interested, struggle to continue and anxiously) before and after click are not obvious; 2) The difference of curve E (viz., be wearisome) before and after click is big.

The characteristics illustrate that the mood swings of users normally occur after going deep into the details of feedbacks but not in the process of browsing the result list. Thus why will the details make users have the obvious mood swings? To answer the question, we exclusively extracted the feedbacks which caused the swings and inspected the probabilities of wrongness at different i-th click (see Table 8). It can be found that nearly all the feedbacks are wrong. Therefore, the mood swing should attribute to the frequently-occurring wrong information, especially that involving obvious errors or/and noises (others normally cannot move the users because their factualness are actually unknown to users at the beginning), e.g. advertisement, dead link, fraud page, irrelevant page, etc.

Table 8. Wrongness of the feedbacks causing mood swings.

i-th click	*1st click*	*2nd click*	*3rd click*
Wrong feedback	1	0.9126	0.9219

4.6 Factualness identification (Q10+Q11)

According to the quantitative evidences, we can conclude the main advantages of highly ranking factual feedbacks: 1) Help users cognize factual knowledge quickly; 2) Avoid to puzzle users in the process of surveying factual knowledge; 3) Reduce the probability of negative mood swings occurring. Therefore, compare to the satisfactory feedback, the factual one can not only supplies trustworthy information service but also improve the user experiments.

Table 9: Success rate of feature in factualness identifying.

Features	*A*	*B*	*C*	*D*	*E*
Factual	0.3439	0.3966	0.1304	0.0949	0.0343
Wrong	0.0796	0.3628	0.0531	0.0620	0.4425
Difference	**0.2643**	**0.0338**	**0.0773**	**0.0329**	**0.4082**

(**A**: *High authority*; **B**: *Credible source*; **C**: *Credible contents*; **D**: *Seems reasonable*; **E**: *I have the same thought*)

Table 10: Success rate of feature in satisfactoriness identifying.

Features	*A*	*B*	*C*	*D*	*E*
Satisfactory	0.2992	0.3333	0.1197	0.0855	0.0086
Dissatisfactory	0.1089	0.2657	0.2352	0.1237	0.0549
Difference	**0.1903**	**0.0676**	**0.1155**	**0.0382**	**0.0463**

(**A**: *Provide everything I need*; **B**: *I can learn some knowledge*; **C**: *Interesting*; **D**: *Beautiful webpage*; **E**: *others*)

Thus how can we identify the factual feedbacks? To answer this question, we inspected the opinions of volunteers on 5 features (see Q10 in Table 2): authority, credibility of source, credibility of contents, reasonability and the cognition consistency. We respectively calculated the probability of the features in being successfully used to determine the factualness (see Table 9). According to the difference between the success rates in factual and wrong feedbacks, we find the authority (A) and cognition consistency (E) should be the effective features.

They both have obvious difference between the success rates. However it is hard to automatically measure the cognition consistency. So the authority should be a feasible feature to identify factual feedbacks. Additionally, by checking the cognition consistency (viz. the users agree with the opinions in the feedbacks), we found its success rate is very low to identify factual feedback. It illustrates the users seldom agree with the opinions in the factual feedbacks. This further proved our hypothesis, users normally cannot precisely determine the factualness when they lack of enough priori or/and relevant knowledge. Besides, we also used the same method to detect the effect features in satisfactoriness identifying. The distributions of success rates are shown in Table 10. It can be found that the feature “providing everything users need” (A) is optimal. Obviously, the users indeed concern whether and how many information is consistent with their preferences, but lack of the consideration to factualness.

4.7 Evaluation Metric and Performance of Traditional IR

In this paper, we propose several new methods to evaluate the performance of information retrieval system in detecting and ranking the factual information. Based on the influences of factual feedbacks to the performance of information retrieval concluded in section 3.2, the factualness-focused evaluation metric should include three quantities to be measured: 1) *P@n*: indicates the precision of the top n feedbacks, which actually is the percentage of the factual feedbacks in the top n feedbacks; 2) NDCG@n[9][10][11]: is the normalized DCG@n which involves the tradeoff between the factualness degree and the ranking; 3) MCost@n: indicate the cost of negative mood swings in the top n feedbacks.

In information retrieval, the traditional DCG comprehensively considers the relevance level and the ranking: if a feedback which has higher relevance level is ranked highly, the DCG will increase, else will reduce. Here, we use the DCG to measure the tradeoff of the factualness level and the ranking: if a feedback which has higher factualness level is ranked highly, the DCG will increase, else will reduce. The third quantity, viz. MCost, is calculated like the DCG but focuses on measuring the tradeoff of the satisfactoriness level of wrong feedback and the ranking: if a wrong feedback which has higher satisfactoriness level is ranked highly, the MCost will increase, else will reduce.

We reproduced two personal information retrieval systems: one used Eye-tracking based user behavior analyzer[1] (abbr., E-sys); the other used click-through based user behavior analyzer[2] (C-sys). And we ran the two systems and evaluate their NDCG@10 and MCost@10 respectively. Besides, we used the metrics to measure the performance of manually labeling results and regarded that as baseline. The performances are shown in Table 11. It can be found that, compared to the baseline, the personal information retrieval systems E-sys and C-sys achieve better NDCG but worse Mcost (higher MCost score means worse user experience).

Table 11. Main performances

Features	*E-sys*	*C-sys*	*Baseline*
NDCG@10	0.2629	0.2816	0.1756
MCost@10	0.3521	0.3343	0.2212

5 Conclusion

This paper gives plenty of evidences achieved from a Google based labeling platform to prove the hypothesis that the factual feedback should be optimal in the search results. In future, we will firstly promote our search-engine based labeling platform. Current version only regards the Wikipedia webpage as the sole factual feedback. But there are indeed some other web pages involving the factual information, e.g. the Baidupedia pages, pages of knowledge bases, etc. Additionally, we will further survey the possible features for factualness identification. In our imagination, the logicality of information should be an effective feature. The logicality can be used to verify whether the relation of events, attribute of things, characteristics of things collaborating in information accord with the factual regularity. This should be a hard but necessary task in our works.

References

Laura, A. G., Thorsten, J., Geri, G. 2004. Eye-Tracking Analysis of User Behavior in WWW Search. The Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM Press, 478-479.

Xue, G-R., Zeng, H-J., Zheng, C., Yong, Y., Ma, W. Y., Xi, W-S., Fan, W-G. 2007. Optimizing Web Search Using Web Click-Through Data. CIKM'04. New York: ACM Press, 118-126.