An Evaluation Method of Web Search Engines based on Users' Sense

Takashi OHTSUKA † Koji EGUCHI ‡ Hayato YAMANA †

† Graduate School of Science and Engineering, Waseda University
3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 Japan
t-ohtsuka@toki.waseda.jp,yamana@yama.info.waseda.ac.jp

‡ National Institute of Informatics
2-1-2 Hitotsubashi Chiyoda-ku Tokyo, 101-8430 JAPAN
eguchi@nii.ac.jp

Abstract

It is not easy to obtain the useful information from the Web space because the space is increasing and large. Therefore, the Web search system becomes indispensable and the improvement of its performance is required. Then, the researches on search engine's evaluation methods are being important. The evaluation based on the conventional evaluation methods, however, is not always equal to the user's evaluation. The reason is most of the conventional evaluation methods are based on precision and recall, which do not take users' sense for search engines into consideration. In this paper, we propose "a user oriented evaluation criterion" that evaluates the performance of Web search systems by considering users ' actions when they retrieve Web pages. We also evaluate the proposed criterion in comparison with the conventional methods by measuring the time spent on search as the users ' satisfaction degree.

Keywords: *Search Engine, Information Retrieval, Evaluation Method.*

1 Introduction

Total capacity of Web space is increasing because the World Wide Web is becoming wide-spread. Therefore, the Web search system is becoming more important to find the information a user wants.

Web documents are unit of information on the Web and information retrieval that mainly related to Web documents is called Web retrieval. Web documents have characteristics that are different from the ones of newspaper articles and theses, which are dealt with conventional information retrieval. To be concrete, Web documents have variability of genres (thesis, catalogue, diary etc. are coexisted), variability of expression (layout used for tag, table and image), and reference of links (reference from page to page by hyperlink), etc. To support these characteristics various methods were suggested and applied for search engines in Web search systems, however we still have many research problems yet.

In the problem of evaluation for Web retrieval in particular, we have lack of evaluation criterion that is appropriate to the Web retrieval. This is caused to insufficiency of studies on evaluation models related to a user's query and its retrieval result. To be concrete, relevance judgment evaluates only texts or additionally considers images etc., and whether or not permits to refer linked pages in retrieval result. How do we evaluate content reliability and its importance? In these ways, conventional evaluation measures of information retrieval exclude many subjects.

In this paper, we propose a new evaluation measure based on Web retrieval characteristics. Concretely, we bring users ' characteristics in case of Web retrieval to evaluation criterion for search systems. Moreover, we consider an application of our approach to clarify points to be improved in Web retrieval systems.

We explain conventional evaluation methods and its characteristics, and describe evaluation criterion that is needed for evaluation of Web retrieval system in the section 2. We propose the new evaluation measure to evaluate retrieval systems by taking into consideration characteristics of Web retrieval and using users ' sense for evaluation criterion in the section 3. In the section 4, we carry out an evaluation experiment to compare conventional evaluation methods with the proposed method. We discuss experimental results in the section 5. Finally, we describe our conclusions in the section 6.

2 Related Work

In this section, we describe the characteristics of evaluation methods that were conventionally used in information retrieval. Then, we explain a new evaluation measure that is designed to cope with modern

information retrieval systems.

2.1 Recall and Precision

Recall and precision are conventional evaluation methods that are designed by Cranfield tests[1] in the middle of the 1950s. We assume a document sets and a query set. In this case, we assume that when the query is given, the number of total relevant documents R can be determined in the document set. Here, we suppose that we perform searching by a retrieval system, and that the retrieval system gets n documents and r relevant documents. In this case, precision is defined r/n, and recall is defined r/R. When recall and precision are used to evaluate for Web retrieval system, we have the following problems;

- Calculation of recall is difficult: we need the total number of relevance documents to get recall but we don't know the total number of relevance documents because database size is too enormous to judge the relevance judgment by human hands. Furthermore we can't calculate actual recall because we can't collect all data on the Web for real retrieval engines.
- Multi-grade relevance judgment is not done because evaluation criterion is binary by relevance or irrelevance: it is difficult to assign Web documents to relevance or irrelevance because Web documents have various genres and expression methods, and binary evaluation is not available.
- Relevance judgment must be done in document sets: recall and precision need test collection that has three elements. Those are document sets that are retrieval object, query sets, relevance document sets for each query. It is difficult to apply test collection that is made artificially to evaluation method for real Web environment.

We have following transformed evaluation measures based on recall and precision to compensate previous problems[2].

- precision(λ): precision for document sets of retrieval result top λ
- recall(λ): recall for document sets of retrieval result top λ

These are measures for evaluating retrieval systems using range of upper rank of retrieval result. We consider that these are suitable for Web retrieval, hoever we don't have clear method to decide the value of λ. If we can't decide the appropriate λ for each query, that may have a great influence on evaluation of system. Furthermore, recall of only retrieval result of top λ has problem of availability when λ is small, because Web retrieval is expected to be applied to the large scale document sets that may have a large number of relevant documents.

- R-precision: precision for document sets of retrieval result top R(R is the total number of relevant documents)

R-precision is a measure when the value of λ is given as the total number of relevant documents at the precision(λ). This can be a measure that indicates effectiveness of retrieval result of upper rank in using test collection situation. However, it is problematic that we generally don't know the total number of relevant documents in Web retrieval.

- (non-interpolated)average precision: from the ranking top calculate precision at the relevant documents appear and average each precision
- n-point averaged precision[3]: average precision at n-point recall that is decided beforehand

We often use eleven points of recall, $0.0, 0.1, \cdots, 0, 9, 1.0$, for n when using the n-point averaged precision.[1] These measures add to evaluation range of lower rank of retrieval result than R-precision. That is to say, these measures evaluate system more macroscopicly. It is difficult to apply these measures for Web retrieval because they are premised on a test collection.

In addition to the above mentioned measure, other measures that based on both recall and precision exists. However,we consider that evaluation measures based on recall and precision are less proper for evaluation of Web retrieval systems because we can't precisely calculate recall in Web retrieval.

In the next section, we describe evaluation measure that is designed to make up for these disadvantages.

2.2 DCG[4]

DCG(Discounted Cumulative Gain) is a recent evaluation measure designed by Järvelin, K. and Kekäläinen, J. in 2000.

DCG can evaluate documents using relevance judgment based on non-binary. Therefore, we can use DCG as the measure of the evaluation criterion with multi-grade relevance. Furthermore, it can also evaluate ranking by considering ranks of relevant documents. We describe the DCG as the following;

We suppose that $d(i)$ indicates i-th-ranked document, $g(i)$ indicates the score of $d(i)$, and $dcg(i)$ indicates the cumulative gain of document's score from top to until i-th-ranked. Then, DCG is defined as follows;

$$dcg(i) = \begin{cases} g(1) & if\ i = 1 \\ dcg(i-1) + g(i)/\log_c(i) & otherwise \end{cases} \quad (1)$$

$$g(i) = \begin{cases} h & if\ d(i) \in H \\ a & if\ d(i) \in A \\ b & if\ d(i) \in B \end{cases} \quad (2)$$

[1] This measure is also referred to as interpolated average precision, since interpolation is usually needed to determine the precision at every point of recall.

where $H, A,$ and B indicate the sets of highly relevant, fairly relevant, and partially relevant documents, respectively. In this way, we can use multi-grade relevance judgment. The value of c, base of logarithm, is weighting factor related to order of rank. If we assume $c > 1$, the larger value of c is, the higher score for a lower ranked document is assigned.

DCG evaluates relevance degrees as well as ranks of relevant documents. If relevant documents appear at upper ranks, the score of the documents is higher; and the lower relevant documents appear, the less score of the documents is applied. DCG evaluates the retrieval system by cumulating score as above.

DCG can evaluate only the range of upper rank and doesn't require the total number of relevant documents. From these points of view, DCG is more suitable measure to evaluate Web retrieval. The problem is that there are not appropriate ways to decide the value of the weighting factors. In addition, DCG requires a test collection because DCG is one of the system-oriented evaluation measures that mainly aim at improvement of performance of retrieval systems. Therefore, we consider that it is not easy for DCG to apply for actual Web retrieval as a evaluation measure. Furthermore, we consider that we are not easy to understand why resulting DCG scores for a system are different from the ones for other systems.

2.3 WRR[5]

WRR(Weighted Reciprocal Rank) is a recent evaluation criterion that was adopted at NTCIR(NII/NACSIS Test Collection for Information Retrieval)[6], which is a retrieval experiment project in Japan started in 1998. NTCIR has referred to retrieval experiments at TREC(Text REtrieval Conference)[7] and conducted four times experiments so far. Web retrieval experiment was conducted at NTCIR-3 for the first time in NTCIR projects. At this experiment, WRR was designed to evaluate Web retrieval in 2001. Web retrieval is also investigated at NTCIR-4 WEB task. We describe the WRR as the following.

We suppose that $d(i)$ indicates i-th-ranked document like in section 2.2. $H, A,$ and B are also defined in the same manner as section 2.2. m indicates top m of a retrieval result of the evaluated target. Then, WRR is defined as follows;

$$wrr(m) = max(r(i)) \tag{3}$$

$$r(i) = \begin{cases} \delta_h/(i-1/\beta_h) & if\ (d(i) \in H \wedge 1 \leq i \leq m) \\ \delta_a/(i-1/\beta_a) & if\ (d(i) \in A \wedge 1 \leq i \leq m) \\ \delta_b/(i-1/\beta_b) & if\ (d(i) \in B \wedge 1 \leq i \leq m) \\ 0 & otherwise \end{cases} \tag{4}$$

where $\delta_h \in \{0,1\}, \delta_a \in \{0,1\}, \delta_b \in \{0,1\}, \beta_b \geq \beta_a \geq \beta_h > 1$ are weighting factors that is satisfied each coefficient.

They have $5, 10, 15, 20$ as value of m in NTCIR evaluation. And they use next two levels as couples of δ_x and β_x.

$$level1 : (\delta_h, \delta_a, \delta_b) = (1,1,0), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$$
$$level2 : (\delta_h, \delta_a, \delta_b) = (1,1,1), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$$

WRR is a measure that evaluates how upper rank a relevant document appears. We evaluate retrieval results by averaging WRR over several search requests. WRR is one of the user-oriented evaluation measures that a user decides whether or not the retrieval system fulfills the user's requirement, because WRR can evaluate only upper range of retrieval results.

However, WRR is apt to binary evaluation whether or not a relevant document appears, because WRR uses extremely upper range of retrieval result. Moreover, it is not easy to adjust the weighting factors in WRR.

3 An evaluation method based on Users ' sense

As mentioned in section 2, researchers mainly use measures based on recall and precision, DCG, and WRR that was used at NTCIR-3 as one of measures to evaluate performance of Web retrieval systems. However, evaluation measures based on recall and precision do not sufficiently take into consideration characteristics of Web documents that are the target document sets of retrieval. It cannot be said that DCG is meaningful if we don't have a large scale test collection that is comparable to the real Web environment. Furthermore, it is not easy for WRR to evaluate with propriety because the parameter setting is complicated.

Therefore, we need new evaluation criterion that can evaluate performance of retrieval systems for enormous Web space. In this section, we describe the characteristics of Web retrieval and propose an evaluation measure based on users ' sense that considers users ' action in performing Web retrieval.

3.1 Characteristics of Web retrieval

When users retrieve in the Web space actually, they usually refer to only about top 30 of retrieval result [8]. That is to say, if a retrieved document is ranked under the 30th order, it can be said to be worthless as retrieval result. Therefore, we evaluate only upper ranked documents as evaluation of retrieval result.

When users refer to a retrieval result, they usually refer to the result from the top by descending order [9]. Therefore, it is desirable for users that the relevant documents are next to each otherin upper range of ranking because it is easy for users to refer to. In this case, we can consider that users 'satisfaction is higher than the case when relevant documents and irrelevant documents are ranked alternately. In other words, it is desirable for the relevant documents to be ranked next to each other, and for the irrelevant documents to be ranked next to each other. For example, if relevance

documents appear in succession, users can judge effectiveness of their query. Furthermore, it leads to reconsider a better query because they can refer to the contents of relevant pages. If irrelevant documents are ranked next to each other, users can change the query immediately.

We suppose the following evaluation measure on the basis of the above mentioned idea.

3.2 Proposal measure

We propose a new evaluation measure UCS (User's Character Score) for adding Web retrieval features to evaluation measure. We define UCS as the following.

We suppose that $d(i)$ indicates i-th-ranked document like in section 2.2 and $s(i)$ indicates score of $d(i)$. Here, we define $s(1) = 1$. Furthermore, we define m like in section 2.3. When X indicates relevant documents sets, we define the UCS score as follows:

$$UCS(m) = \sum_{i=1}^{m} s(i) \tag{5}$$

$$s(i) = \begin{cases} 1 & if\ (d(i-1) \in X \wedge d(i) \notin X) \\ 1 & if\ (d(i) \in X \wedge d(i-1) \notin X) \\ a \cdot s(i-1) & if\ (d(i-1), d(i) \in X) \\ a \cdot s(i-1) & if\ (d(i-1), d(i) \notin X) \end{cases} \tag{6}$$

where a is a weighting factor that satisfy $a > 1$.

When relevance documents or irrelevance documents are ordered sequentially in retrieval result, each document is given high score in the equation of UCS. In other words, score of $d(i+1)$ is higher when relevance documents (or irrelevance documents) are ranked continually. Finally, UCS can be obtained as the sum of $s(i)$ at upper m retrieval result. UCS can evaluate only upper retrieval result that is referred to users by setting the value of m is about 30. That is to say, we give the same level of worth for the documents that are expected to be referred by users, and we consider as worthless for the documents that are expected not to be referred by users. Furthermore, each documents score is influenced by only continuity of ranking of relevant documents or irrelevant ones, but not influenced by ranking. We focused on document order because document order is important factor in browsing retrieved documents. Therefore, we can consider that UCS is a measure of users 'satisfaction degree as mentioned in section 3.1.

3.3 Proposal evaluation method

We propose the following evaluation method, considering users 'sense.

We evaluate time spent on search as evaluation criterion from a different viewpoint from UCS. The spent time means the time from start of a user 's referring to retrieval result until the user judge the retrieval result as satisfied or dissatisfied. Then we request each user to judge whether the retrieval result is satisfied or dissatisfied. By the following two reasons we introduced the binary judgment of satisfaction or dissatisfaction.

- Judgment criterion setting is difficult to be multi-grade (As the result, this can moderate influence by individual difference)
- To avoid evaluation judgment to be incline to intermediate between satisfaction and dissatisfaction

We evaluate in the two directions where we give low evaluation when spent time is long or retrieval result is judged as dissatisfied. In other words, we give low evaluation to retrieval result when the searching time is long and it is judged as dissatisfied, and give high evaluation to retrieval result when the searching time is short and it is judged as satisfied.

We use time spent on search, judgments on retrieval result as satisfied or dissatisfied and UCS as evaluation methods for IR systems.

4 Evaluation experiment

We carry out an experiment to evaluate characteristics of the proposed measure. In this section, first of all, we describe the data sets that are used in our experiment. Secondly, we describe details of the experiment, and then show the result of the experiment.

4.1 Data sets

We describe the data sets of this experiment.

4.1.1 An object of retrieval

As the retrieval target, we used the data sets constructed for the NTCIR-3 Web Retrieval Task (NW100G-01), which were collected from ".jp " domain Web servers from August to November in 2001. The total volume is about 100 gigabytes, in which each document consists of a text file and its meta data.

We used five independent retrieval results that were submitted to NTCIR-4 WEB Task by independent participating teams. Retrieval result is included in the document sets used for NTCIR-3 Web Retrieval Task (NW100G-01) because NTCIR-4 WEB Task used the NW100G-01 data as the retrieval target.

4.1.2 Topics

As the topics, we used the topic data [10] that was used at NTCIR-4 WEB Task (the Informational Subtask) [11]. To be concrete, 267 topic candidates were made by 20 persons. When they made the topics in 2003, they exclude topics that don't exist in those days because the retrieval target data were gathered in 2001. Organizers of NTCIR-4 WEB Task selected 153 topics by excluding unsuitable ones. After NTCIR-4 participating teams submitted their retrieval result, the organizers divided the topics into two groups by analyzing those results. One is (a) for evaluation based on exhaustive relevance judgment, and the other is (b) for evaluation based on relevance judgment that evaluates

only upper retrieval result. The above-mentioned (a) has about 50 topics and (b) has about 100-150 topics (including topics of (a)). We used 53 topics that were extracted from topics of (a) by removing topics that has no relevant documents, few relevant documents or extremely many relevant documents like several hundred thousand in this experiment. Figure 1 indicates a topic example.

<TOPIC>
<NUM>0002</NUM>
<TITLE CASE="b"> トランペット, 価格, 特徴 </TITLE>
<DESC> トランペットの特徴およびその価格が記述されている文書を探したい。</DESC>
<NARR>
<BACK> 店舗ごとにどのような特徴をしてどれ位の値段のトランペットが売られているのかを知りたい。</BACK>
<TERM>「トランペット」とは、真鍮製でオーケストラや吹奏楽、ジャズの演奏会などで一般的に使用される種類の楽器を指す。プラスチック製などの、玩具としてその名が付いたものは該当しない。</TERM>
<RELE> トランペットの製品番号・その形状や音色などの特徴・店頭価格が全て記載されている文書が適合する。</RELE>
</NARR>
<ALT0 CASE="b"> トランペット </ALT0>
<ALT1 CASE="b"> トランペット, 特徴, 価格 </ALT1>
<ALT2 CASE="b"> トランペット, 特徴, 値段 </ALT2>
<ALT3 CASE="c" RELAT="2-3"> トランペット, 特徴, 価格 </ALT3>
<USER> 大学 2 年, 女性, 検索歴 6 年, 熟練度 3, 精通度 4</USER>
</TOPIC>

Figure 1. Topic example

Each topic is composed of the following items.

- <NUM> ('topic number') indicates topic ID number.
- <TITLE> ('title') is 1-3 query terms that a topic creator assumed to input to real search engines. These are listed in the order of importance for searching. <TITLE> has "case ' ' attribute that indicates one of the following retrieval strategy types.
  - The case when we can use OR operator for the relation between all terms
  - The case when we can use AND operator for the relation between all terms
  - The case when we can use OR operator for the relation between only two terms out of three terms.
    - Those two terms are specified by "RELAT ' ' attribute.
- <DESK> ('description') is the most fundamental description of information needs and expressed in about one sentence.
- <NARR> ('narrative') describes background and purpose of retrieval, definition of the terms, and supplement of relevance judgment criterion in several paragraphs. These are indicated by <BACK>, <TERM> and <RELE> tags, respectively. <BACK> is always described, but the other two tags can be omitted.
- <ALT0> ('alternative query 0') is one query term that was extracted from the top of TITLE. However, this was omitted when TITLE is expressed in one term. TITLE is 1-3 query terms that the topic creator assumed to input to real search engines. These are listed in the order of importance for searching. Therefore, ALT0 is defined by the most important term for searching.
- <ALT1>, <ALT2>, <ALT3> ('alternative query 1, 2 and 3') is query that was originally given by three persons who were different from the topic creator when he/she referred to topic (at that time, TITLE tag was deleted). However, all tags by three persons were not defined because the tag that is the same as TITLE was omitted. ALTn format is equivalent to TITLE except for the tag name.
- <USER> ('user attributes') provides the attributes of the topic creator, such as the job title, gender, search experience, level of search skill, and level of familiarity with topics.

4.2 Explanation of experiment

We experimented to verify the consistency between the proposed evaluation method and real users 'sense.

We showed evaluators the five kinds of retrieval results that were given by IR systems of independent participating teams at NTCIR-4 WEB Task (the Informational Subtask). The display order of five kinds of retrieval results was changed because we needed to moderate influence of time spent on search by display order or influence of retrieval evaluation.

We showed the evaluators a retrieval result after they understood each topic. We showed upper 30 retrieval result that was assumed to be referred by real users. The evaluators evaluated each retrieval result shown. The evaluator are 26 persons who often use the Internet.

The evaluators evaluated satisfaction or dissatisfaction of retrieval result by referring to shown retrieval result actually. However, the evaluators did not need to refer entire retrieval result. The evaluators could make a prompt decision of satisfaction or dissatisfaction using only a part of the retrieval result. This can evaluate the degree of satisfaction on the basis of document continuity, and so we consider that this can evaluate retrieval process and user interaction.

4.3 Evaluation methods

We use the proposed evaluation method for evaluating retrieval result.

We compare UCS with DCG and WRR to carry out comprehensive user-oriented evaluation. We set the each parameter of evaluation measure as the following; To calculate DCG, each parameter was set as $(h, a, b) = (3, 2, 1)$ and $c = 2$ in section 2.2. To calculate WRR, each parameter was set as $(\delta_h, \delta_a, \delta_b) = (1, 1, 1), (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$ in section 2.3. To calculate UCS, parameter a was set as $a = 1.1$ in section 3.2.

4.4 Results of Experiment

First of all, we show in table 1 the UCS, DCG, WRR and average time spent on retrieval over 53 topics for five kinds of retrieval results. We give two kinds of scores id_lack and id_poss for scoring DCG and UCS in table 1, where id_lack treats duplicate pages as irrelevant. Moreover, id_poss doesn't consider duplicate pages, and so depends only on each document ' s relevance degree.

Table 1. Comparison of UCS, DCG and WRR with average time spent on search

	DCG	DCG	WRR	UCS	UCS	avg.spent time
	id_lack	id_poss		id_lack	id_poss	
res_1	7.70	16.39	0.67	47.66	50.51	74.09
res_2	4.97	11.09	0.44	56.89	56.20	90.42
res_3	3.71	7.59	0.38	76.60	76.62	79.15
res_4	1.45	3.38	0.17	111.93	113.60	80.94
res_5	7.33	15.73	0.64	47.73	54.32	70.30
ave.	5.03	10.84	0.46	68.16	70.25	78.98
cor.coef.	-0.51	-0.49	-0.54	0.28	0.20	

We find from table 1 that average time spent on search is shortest in retrieval result 5, and 8.6 seconds shorter than total average time spent on search. Therefore, we can judge retrieval result 5 is the most understandable. Contrary, we can find retrieval result 2 is the least understandable because retrieval result 2 has 11.4 seconds longer average time spent on search. We calculated correlation coefficient between score of each measure and average time spent on search, however we could not find significant correlation.

Secondly, we show in table 2 the UCS, DCG, WRR and average satisfaction score for the five kinds of retrieval results, where average satisfaction score is the average score of satisfaction score (perfect score is 26 points) that is the number of evaluators who judged as satisfied for each topics. We find in table 2 that retrieval result 1 is the best retrieval result because the average satisfaction score is the highest. Contrary, we can judge that retrieval result 4 is the worst retrieval because the average satisfaction score is the lowest. In the same way as in table 1, we calculated correlation between coefficient of score of each measure and average satisfaction score. We find id_lack and id_poss in DCG and WRR have high correlation with average satisfaction score. Moreover, we find id_lack and id_poss in UCS have high negative correlation with average satisfaction.

Table 2. Comparison UCS, DCG and WRR with average satisfaction score

	DCG	DCG	WRR	UCS	UCS	ave.sat. score
	id_lack	id_poss		id_lack	id_poss	
res_1	7.70	16.39	0.67	47.66	50.51	17.81
res_2	4.97	11.09	0.44	56.89	56.20	12.17
res_3	3.71	7.59	0.38	76.60	76.62	7.58
res_4	1.45	3.38	0.17	111.93	113.60	3.72
res_5	7.33	15.73	0.64	47.73	54.32	17.08
ave.	5.03	10.84	0.46	68.16	70.25	11.67
cor.coef.	0.99	1.00	0.98	-0.95	-0.92	

5 Discussion

The score of UCS tend to be high when irrelevant documents ranked continuously, in comparison with DCG or average satisfaction degree. Therefore, we can't directly compare UCS with DCG or WRR. Therefore, we propose UCS2 that can directly compare with by changing a parameter as another measure.

UCS2 gives different score for continuously ranked irrelevant documents from continuously ranked relevant ones. To be concrete, we changed the parameter a as $a = 0.9$ in section 3.2 (we used $a = 1.1$ for continuously ranked relevant documents). This means that continuously ranked irrelevant documents take a penalty. We show in table 3 the comparison of UCS2 with DCG and WRR. Correlation coefficient with UCS2 is calculated like the following in table 3. We calculated correlation between id_lack of UCS2 and id_lack of DCG, and between id_poss of UCS2 and id_poss of DCG. In the same way, we calculated correlation coefficient for id_lack and id_poss of WRR.

Table 3. Comparison UCS2 with DCG,WRR

	DCG	DCG	WRR	WRR	UCS2	UCS2
	id_lack	id_poss	id_lack	id_poss	id_lack	id_poss
res_1	7.70	16.39	0.67	0.67	27.99	31.83
res_2	4.97	11.09	0.44	0.44	23.59	25.08
res_3	3.71	7.59	0.38	0.38	21.23	24.04
res_4	1.45	3.38	0.17	0.17	18.90	21.42
res_5	7.33	15.73	0.64	0.64	27.18	35.14
ave.	5.03	10.84	0.46	0.46	23.78	27.50
cor.coef.	0.99	0.93	0.99	0.93		

We find in table 3 that UCS2 has extremely high correlation DCG and WRR.

Form these result, we can consider that we can use UCS2 as substitute of DCG. We can consider that relevance degree is related to document continuity because correlation between UCS2 and DCG is high. We consider the following characteristics of UCS2 as advantage.

- UCS2 can evaluate IR systems at few costs using only binary judgment (relevant or irrelevant) for upper m ranking.
- UCS2 doesn't need large scale test collection that is needed by DCG, or multi-grade relevance judgment that is difficult to create the criterion for judgment.

We show in table 4 that verification result of correlation between UCS2 and average time spent on search. We find in table 4 that correlation between id_poss

Table 4. Comparison UCS2 with average time spent on search

	UCS2	UCS2	avg.spent time
	id_lack	id_poss	
res_1	27.99	31.83	74.09
res_2	23.59	25.08	90.42
res_3	21.23	24.04	79.15
res_4	18.90	21.42	80.94
res_5	27.18	35.14	70.30
ave.	23.78	27.50	78.98
cor.coef.	-0.53	-0.73	

and average time spent on search becomes stronger, however, correlation between id_lack and average time spent on search is at the same level of the ones for DCG or WRR in comparison with table 1. We can consider that this feature comes from user understandability since UCS is based on continuity. We deemed duplicated documents as irrelevant for id_lack. Therefore, we can suppose that correlation is at the same level for DCG or WRR because influence of document continuity becomes weaker.

6 Conclusion

As future issues, we need to design more interactive evaluation methods that can evaluate not only retrieval result given by a one-shot query but also retrieval results given by continuous queries.

In this paper, we proposed an evaluation method for web search engines on the basis of users ' sense. We couldn't find correlation between the proposed measure and time spent on search from experimental results. We found correlation between user satisfaction score and DCG by changing a parameter of the proposed measure. Thereby, we could bring users ' sense in evaluation. We think that the proposed measure can apply evaluation of real IR system because the proposed measure can be expected to reduce costs of creating test collections.

Acknowledgment

This research was partially supported by the Expenses for Promotion of Science and Technology "e-Society" and the 21st Century COE Program "Productive ICT Academia" from the Ministry of Education, Culture, Sports, Science and Technology. In this research, we used a part of data of NTCIR-4 WEB task that is sponsored by National Institute of Informatics, as organizers of NTCIR-4 WEB task project.

References

[1] C.W. Cleverdon, "The significance of the Cranfield tests on index languages, " In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12, ACM Press, Chicago, October 1991.

[2] David Hawking, Ellen Voorhees, Nick Craswell, Peter Bailey, "Overview of the TREC-8 Web Track, "Proceedings of the 8th Text REtrieval Conference, NIST Special Publication 500–246, pp.131–149, 1999.

[3] R.Baeza-Yates, "Modern Information Retrieval, "Addison Wesley Longman Publishing, 1999.

[4] Kalervo Järvelin & Jaana Kekäläinen, "IR evaluation methods for retrieving highly relevant documents, "Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.41–48, New York, 2000.

[5] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama, "Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure, "IEICE Transactions on Information and Systems, Vol.E86-D, No.9, pp.1804–1813 , Sep. 2003.

[6] "NII-Test Collection for IR Home Page," $\langle http : //research.nii.ac.jp/ntcir/ \rangle$.

[7] "Text REtrieval Conference (TREC) Home Page," $\langle http : //trec.nist.gov/ \rangle$.

[8] Amanda Spink, B. J. Jansen, D. Wolfram & T. Saracevic, "From E-Sex to E-Commerce: Web Search Changes," IEEE Computer, 35(3), pp.107–109, Pennsylvania, Mar.2002.

[9] "Japan internet.com," $\langle http : //japan.internet.com/research/20020320/1.html \rangle$.

[10] Koji Eguchi, et al., "Overview of the Informational Retrieval Task at NTCIR-4 WEB," Working Notes of the 4th NTCIR Workshop Meeting, Tokyo, June 2004 (to appear).

[11] "NTCIR-WEB," $\langle http : //research.nii.ac.jp/ntcweb/ \rangle$.