Analysis of Korean Patent & Trademark Retrieval Query Log to Improve Retrieval and Query Reformulation Efficiency(

질의로그 데이터에 기반한 특허 및 상표검색 에 관한 연구

Jee Yeon Lee(이지연)((

Woojin Paik(백우진)(((

Abstract

To come up with the recommendations to improve the patent & trademark retrieval efficiency, 100,016 patent & trademark search requests by 17,559 unique users over a period of 193 days were analyzed. By analyzing 2,202 multi-query sessions, where one user issuing two or more queries consecutively, we discovered a number of retrieval efficiency improvements clues. The session analysis result also led to suggestions for new system features to help users reformulating queries. The patent & trademark retrieval users were found to be similar to the typical web users in certain aspects especially in issuing short queries. However, we also found that the patent & trademark retrieval users used Boolean operators more than the typical web search users. By analyzing the multi-query sessions, we found that the users had five intentions in reformulating queries such as paraphrasing, specialization, generalization, alternation, and interruption, which were also used by the web search engine users.

초록

본 연구는 특허 및 상표 검색 개선을 위한 방법을 제안하고자 하는 목적에서 출발하였다. 이를 위해 193일간 한국특허정보원의 특허기술정보서비스를 이용한 17,559명의 이용자가 작성한 100,016개의 질의문에 대한 로그 데이터를 분석하였다. 개별적인 질의로그 분석 이외에, 2,202개의 복수 질의문을 이용한 탐색세션을 분석함으로써 검색 개선을 위한 추가적인 단서를 발견하였다. 분석결과에 의하면, 특허 및 상표검색은 일반적인 웹 검색의 유형과 유사한데, 특히 질의문의 길이가 짧다는 측면에서 매우 흡사하다. 그러나 특허 및 상표검색의 경우, 일반 웹 검색보다 불리언 연산자를 많이 사용하고 있었다. 복수 질의문 분석을 통해 이용자들이 질의문을 재작성하는데 도움이 될 수 있는 탐색기능을 제안할 수 있었다. 복수의 질의문으로 구성된 탐색세션을 분석한 결과, 이용자들은 질의문을 재작성하기 위하여 부연하기, 특정화하기, 일반화하기, 교체하기, 중단하기와 같은 방법을 사용하고 있음을 알 수 있었다.

Keyword: Patent & Trademark Retrieval, Query Log, Query Reformulation Efficiency

1 Introduction

There have been a number of studies on analyzing query logs of web search engines. One of the most significant reports was on the analysis of AltaVista query log (Silverstein et al., 1999), which found patterns and computed the statistics of about 1 billion search requests over a six-week period. In addition, there were numerous studies on the transaction log analysis of Online Public Access Catalog (OPAC), which focused on the user interaction with online catalogs. One such study is by Blecic et al. (1998), which defined transaction as a query by the user followed by a response from the system. Some of the studies provided findings about the patterns discovered in terms of the descriptive statistics or the correlations amongst the query terms. Others provided suggestions to improve the retrieval effectiveness either by altering the original queries or by changing how the queries are entered or the results are shown. However, there were no query log analysis studies on the specialized search engines such as the patent & trademark retrieval systems. Furthermore, there were few research projects focusing on the retrieval efficiency improvement by incorporating the query log analysis.

In each year, 290,000 new patents and trademarks registered in Korea and more than 12 million new patents and trademarks registered in the world. The coverage of the Korean Patent & Trademark Retrieval System is all Korean patents and trademark documents as well as US, Japan, Europe, and International Patents under the Patent Cooperation Treaty (PCT) (Yoo 2004).

The system used to retrieve Korean patent and trademark information is an extended Boolean system, which allows the users to include Boolean, near, and wildcard related operators. The system returns ranked list of patents or trademarks according to the relevancy to the submitted query. The users can also include field search restrictions based on the typical fields embedded in the patent and trademark documents.

We analyzed 100,106 patent & trademark retrieval queries submitted over 193 days to find clues to improve the retrieval efficiency especially to make the retrieval speed faster. We conducted a series of extensive statistical analyses of the query log. We analyzed the query log from various points of view including: 1) the users who submitted the queries; 2) time period when the queries were submitted; 3) the number of query terms; 4) Boolean operators in the queries; 5) field searching restrictions; and finally 6) multi-query sessions. Based on the analysis, we were able to come up with a number of practically usable recommendations to improve the retrieval efficiency.

2 Survey of Previous Works

Altavista query log study (Silvestein et al., 1999) provided extensive information at the session, query, and term level as well as term correlation. Our query log analysis methods were influenced by this study and thus there are numerous methodological similarities except that the Altavista study did not include the query reformulation related explorations and we have not investigated the query term correlation.

There was a study to analyze the Excite (http://www.excite.com) web

search query log (Jansen, et al., 2000). It analyzed 51,473 queries submitted by 18,113 users. The analysis result included the number of queries per user, the number of query terms, the number of retrieved documents viewed by the user, the degree of query complexity, and the number of query modification by relevance feedback. In our study, we were not able to analyze the user actions after the retrieval results are returned as there is no corresponding data in the query log. In addition, our multi-query session analysis did not consider the relevance feedback process into account as there was no retrieval results in response to the initial queries. The limitation of our study is the source data not having the retrieved patents or the trademarks in the query log.

About 16 million queries, which were submitted to the Fireball system (http://www.fireball.de), were analyzed as a part of a large web-searching study (Hoelscher, 1998). The analysis resulted in the average query length to be 1.66 query terms. This is somewhat shorter than the patent & trademark queries. However, the statistics concerning the number of query terms is similar in that 54% of the queries included only one query term and less than 2% of the queries had five or more terms. However, there was significant difference in terms of using Boolean operators. Less than three percent of the Fireball engine users included one or more Boolean operators in the queries. Our analysis of the patent & trademark query log resulted in about 37% of the queries included one or more Boolean operators.

Noriaki et al. (2003) analyzed web search logs to extract semantic relations among the query terms. They claimed that the user■s intentions guide the user to change the terms in the successive queries. They also claimed that the user■s intentions can be used to clarify the semantic relation among words and then eventually enable an ontology or a thesaurus to be developed automatically. They categorized user■s intentions into five classes. They are paraphrasing, specialization, generalization, alternation, and interruption. Paraphrasing is restating the initial query with different terms. Specialization is narrowing the focus of the initial query in the subsequent queries. Generalization is broadening the initial query focus. Interruption is caused by the change of the information needs and thus the subsequent query is not related to the initial query. Alternation is in between generalization and paraphrasing in that it represents the slight shift of the query focus. We have adopted this classification scheme and further developed a set of query reformulation strategies. The user■s intention classification and the query reformulation strategies were used in our multi-query session analysis.

3 Query Log Analysis

The query log analysis that we conducted consists of two categories. The first category is the analysis of individual queries and the second category is the analysis of how queries are modified during the sessions. This type of analysis referred as first-order analysis of queries (Silverstein et al., 1999). In this study, we refer first-order analysis as mainly analyzing the frequencies of the queries, query terms, Boolean operators, and field search requirements.

3.1 The Query Log Data Set

The query log consists of 100,016 queries submitted by 17,559 unique users collected over 193 days from September 26, 2002 to April 6, 2003. This log includes all queries submitted to the patent & trademark retrieval system accessible from Korea Institute of Patent Information (http://www.kipris.or.kr) over this time period.

<Table 1> Sample Queries from the Korean Patent & Trademark Retrieval System Query Log

USERID0001|(감자<and>껍질)<in>(TL,AB)|20021015215238
USERID0002|(한맥)|2002101521
USERID0004|((특허법인아람)<in>(AG))|20021016000654
USERID0004|((BC*<or>매몰*<or>베리드*<or>콘택*<or>컨택*<or>contact*<or>뷰리드*<or>(buried<near/2>contact))<and>((스토리지*<or>저장*<or>storage*<or>하부*<or>바탐*<or>bottom*)<near/2>(전극*<or>폴리*<or>노드*<or>poly*<or>node*))<and>(지그재그*<or>zigzag*))<in>(TL,AB,CL)<and>((H01L*)<in>(IPC))|2002101516
USERID0005|(파이프 보호)<in>(TL,AB)|2002110820
USERID0005|(파이프 마개)<in>(TL,AB)|20021108202403
USERID0006|((엔바이로테크)<in>(AP))|2003011502

The data set is a text file. Its size is 6.57MB. Each query occupies one line in the file. While many web search engine logs, which were used in the previous studies, include search output in response to the queries and the records of what the users did with the results, this patent & trademark query log does not include them. Thus, it was not possible to analyze the user actions after the results came back.

Typical query entry consists of a user id followed by query terms joined by optional Boolean operators. The query terms can be enclosed by parentheses to show the precedence information. The query terms can be constrained by field identifier such that the search scope is limited to the designated fields. The last component of the query entry is the query submission time. The time component consists of four-digit year followed by two-digit month, two-digit-date, two-digit hour, two-digit-minute, and finally two-digit second. However, the time components do not always have the time accuracy up to the seconds. Some queries have time up to day or hour. The components are delimited by ■|■ symbol. The Table 1 shows the seven queries from the log. The actual user ids are replaced with pseudo id to protect the identities of the users.

The first query log entry in the Table 1 means that the user with the id, ■USERID0001■ submitted a query with two terms, which were joined by an ■and■ Boolean operator. The first term in the query was ■감자', which means potato. The second term was '껍질', which means skin. There was a field constraint in the query saying that the Boolean operator joined two terms must be in either the TL (title) field or the AB (abstract) field. Finally, the query was submitted at 38 seconds after 9:52pm on October 15, 2002.

The query log does not have any specific session demarcation. Thus, we made an assumption to identify the sessions. Since the queries in the log are

ordered by the time of submission, we regarded two or more consecutive user ids appearing in the log as a single session. This method of session identification will yield less number of sessions as there can be a user X, who happens to submit a query while another user Y is submitting a series of queries in a row. In this case, a single session by a user Y will be treated as two sessions. However, we decided to operate under this assumption as we wanted to analyze how the queries are modified during a session even under this imprecise assumption. Under this assumption, the fifth and the sixth rows constitute a single session by one user.

3.2 Analysis of Submitted Queries by User

By analyzing the number of queries with respect to the total number of unique users, the results in the Table 2 were computed.

<Table 2> Statistics concerning the Number of Queries Submitted by User

Average number of queries per user	5.7
Median number of queries per user	2.0
Standard Deviation of number of queries per user	12.7
Maximum number of queries per user	616.0
Minimum number of queries per user	1.0

To understand how many queries were submitted by the frequent users, we computed the figures as shown in the Table 3. We define the frequent user as the user, who submits queries comparatively more than other users. The first column in the Table 3 shows the rank of the users by the most frequent user. Thus, the most frequent user, who submitted 616 queries during the 193-day period is ranked as number one. There are 5,857 users, who submitted only one query during the 193-day period. Thus the last row in the column shows the user rank of the 5,857 tied users, who submitted only one query, as 11,703. The second column refers to the cumulative percentage of the user ranking listed from the most frequent to the least frequent as shown in the first column. If a user is ranked as $1{,}755^{th}$ frequent user then the corresponding value for the user in the second column will be 10% by dividing 1,755 by 17,559, which is the total number of unique users. The third column is the number of queries submitted by the corresponding user whose ranking is shown in the first column. The value of the cell, which is located in the intersection of the third column and the second row is 127. This means the user, who is ranked as the 25^{th} frequent user submitted 127 queries over the 193-day period. The fourth column shows the cumulative number of queries submitted. The value of the cell, which is located in the intersection of the fourth column and the second row is 5,024. It refers to the total number of queries submitted by the frequent users, who are ranked from the top to 25^{th}. The fifth column is the cumulative percentage of the number of queries submitted over the total number of queries. This value is computed by dividing the corresponding cumulative number of queries submitted shown in the fourth column divided by 100,016, which is the total number of queries.

<Table 3> Statistics concerning the Cumulative Percentage of Number of Queries Submitted over Total Number of Queries in the Log from the Most Frequent User

User Rank from the Most Frequent User	Cum. Percentage of Users over Total Number of Users from the Most Frequent User (%)	Num. of Queries Submitted	Cum. Num. of Queries Submitted	Cum. Percentage of Number of Queries Submitted over Total Number of Queries (%)
1	0.01	616	616	0.62
25	0.14	127	5024	5.02
78	0.44	79	10053	10.05
150	0.85	60	15009	15.01
244	1.39	48	20037	20.03
365	2.07	37	25013	25.01
521	2.96	29	30009	30.00
714	4.06	23	35020	35.01
952	5.41	19	40019	40.01
1245	7.07	15	45006	45.00
1607	9.13	13	50014	50.01
2043	11.61	10	55009	55.00
2572	14.61	9	60006	60.00
3222	18.31	7	65008	65.00
4008	22.77	6	70008	70.00
4942	28.08	5	75008	75.00
6099	34.66	4	80009	80.00
7605	43.21	3	85011	85.00
9628	54.71	2	90011	90.00
11703	100.00	1	100016	100.00

The Table 3 does not show all possible values as listing all possible values will make the table too big to include. We selected only the representative values based on the 5% increment of the values in the fifth column. The X-axis in the Figure 1 corresponds to the third column in the Table 3. The Y-axis corresponds to the fifth column. The Figure 1 is the result of charting of all values. As one can observe from the Table 3 and the Figure 1, small number of frequent users submitted a large percentage of the queries in the log. For example, 9.13% of the top frequent users accounted for 50.01% of all the queries. This observation led us to the first retrieval efficiency suggestion, which can be described as dedicating one or more servers for the known frequent users. If this suggestion is implemented then the overall retrieval speed can be maximized as we can avoid the frequent users hogging up the retrieval system resources. This improvement will eventually make many more users satisfied with the retrieval efficiency. We refer this suggestion as taking care of the power users.

<Figure 1> Plotting the cumulative percentage of the number of questions submitted over the total number of queries against the number of queries submitted

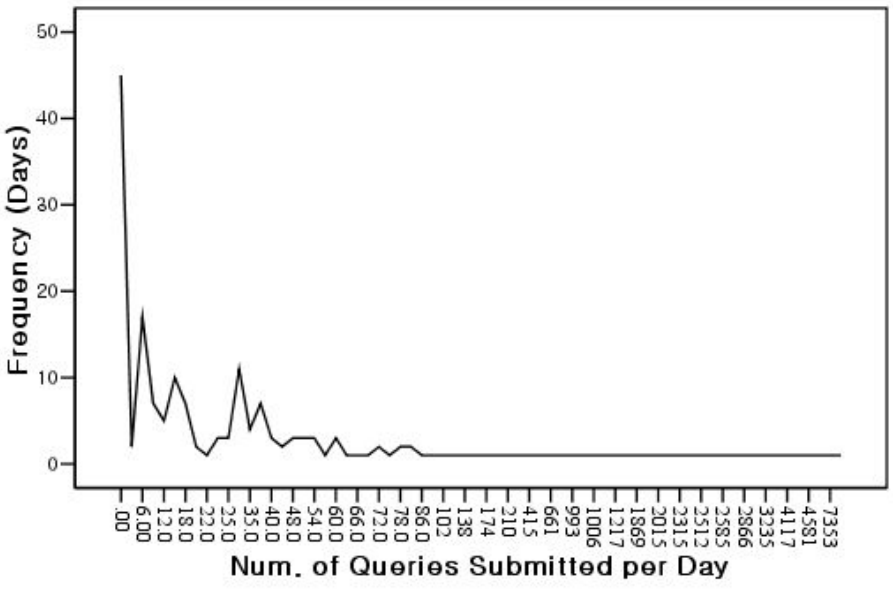
3.3 Analysis of Submitted Queries by Day

By analyzing the number of queries submitted over 193 days, the results in the Table 4 was computed.

<Table 4> Statistics concerning the Number of Queries Submitted by Day

Average number of queries per day	518.1
Median number of queries per day	24.0
Standard Deviation of number of queries per day	1,609.4
Maximum number of queries per day	13,225.0
Minimum number of queries per day	0.0

No queries were submitted for 45 days during the 193-day period. These 45 days probably include the times when the patent & trademark retrieval system was not providing services for whatever reasons and thus the users were not able to submit queries. However, there was no data to identify the actual days when no queries were submitted and the days when the users were not able to submit the queries. Thus, we included all days in the analysis.



<Figure 2> Plotting the number of questions submitted for one day against the number of days

The Figure 2 shows the frequency distribution of the number of queries submitted. Although, there were days when many hundreds or

even thousands of queries were submitted, these days are a few. The Figure 2 shows that the frequency distribution is strongly skewed to the right. It also explains why the standard deviation is unusually large for this statistics. In summary, the median is better than the average to use if one wants to estimate the typical number of queries submitted for one day as median is less sensitive to the extreme outliers.

We can also analyze the number of the queries submitted by days of the week. Our assumption was that it might be possible to assign additional hardware on certain days of the week to maintain constant level of retrieval efficiency if one or more days of the week had comparatively larger number of submitted queries. Table 5 shows the analysis result of the number of queries submitted by the days of the week. By comparing the average number of queries submitted, the users submitted the largest number of queries on Friday and least number of queries on Sunday.

<Table 5> Statistics concerning the Number of Queries Submitted by Days of the Week

	N	Mean	Standard Deviation	Minimum	Maximum
Monday	27	349.63	773.37	0	2770
Tuesday	27	441.30	1472.87	0	7353
Wednesday	27	781.48	2464.35	0	11579
Thursday	28	563.50	1352.53	0	4581
Friday	28	971.54	2610.16	0	13255
Saturday	28	395.18	829.76	0	3235
Sunday	28	124.50	280.60	0	1006
Total	193	518.07	1609.35	0	13255

To determine whether there are significant differences amongst the days of the week in terms of the number of queries submitted, we applied One-way Analysis of Variance procedure. Although, the means look different, it turned out to be that the numbers of submitted queries by the days of week are not significantly different statistically ($F=0.857$, $p=0.528$).

3.4 Analysis of Query Terms

By analyzing the number of query terms, the results in the Table 6 was computed. There were 276 queries, which did not have any valid query terms. They were either empty or had only field specification. The Table 7 shows an abbreviated version of the query term occurrence frequency distribution. The analysis revealed that most of the queries are short. More than 90% of the queries had three or less query terms. About 61% of the queries had only one query term. The queries with seven or more query terms make up less than 1% of the total queries.

<Table 6> Statistics concerning the Number of Query Terms, which Make-up the Queries

Average number of query terms in a query	2.1
Median number of query terms in a query	1
Standard Deviation of number of queries per day	3.1
Maximum number of query terms in a query	103
Minimum number of query terms in a query	0

Since, the patent & trademark retrieval system takes more time to get the results as the number of query terms increases, the overall system efficiency in terms of the retrieval speed will be optimized if there can be multiple servers, which can process queries based on the number of query terms.

<Table 7> Frequency Distribution of the Number of Query Terms

Number of Query Terms	Frequency	Percent	Cumulative Percent
0	276	0.28	0.28
1	61404	61.12	61.39
2	15768	15.77	77.16
3	13056	13.05	90.21
4 or more	9788	9.79	100.00

3.5 Analysis of Boolean Operators in the Queries

36.97% (36,972/100,016) of the queries had at least one of the AND, OR, and NOT operators. The Table 8 shows the statistics of the queries with each Boolean operator.

<Table 8> Statistics concerning the Number of Queries with Specific Boolean Operators

Boolean Operator	Number of Queries with Operator / Total Number of Queries	Percentage
AND	35,091/100,016	35.09
OR	7,995/100,016	7.99
NOT	307/100,016	0.31

More specifically, we analyzed multiples of the Boolean operators co-occur in the queries. The Table 9 shows analysis results.

<Table 9> Statistics concerning the Number of Queries with Specific Combinations of Boolean Operators

3.6 Analysis of Field Search in the Queries

There were 45 field search restrictions imposed on the queries. Out of 45, six field search restrictions were the combination of two or more fields. The combinations of the field restrictions were treated as disjunctions by the patent and trademark retrieval system as the query term restricted by the field restrictions can occur in any one of the

	Frequency of Occurrence	Percentage of Occurrence over Total Num. of Queries	Percentage of Occurrence over Num. of Queries with AND	Percentage of Occurrence over Num. of Queries with OR	Percentage of Occurrence over Num. of Queries with NOT
Queries with both AND and OR	6,177	6.18	17.60	77.26	-
Queries with both AND and NOT	233	0.23	0.66	-	75.90
Queries with both OR and NOT	91	0.09	-	1.14	29.64
Queries with AND, OR and NOT	80	0.08	0.23	1.00	26.06

listed fields to satisfy the query. The rest of the restrictions were made of single field.

The Figure 3 shows the frequency distribution of the top nine field search restrictions. The rest of the field search restrictions were combined into a miscellaneous category. Some of the two-character symbols mean the following: AN for ■Application Number■, AP for ■Applicant Name/Code■, AD for ■Application Date■, TL for ■Title■, AB for ■Abstract■, GN for ■Registration Number■, IN for ■Inventor Name■, TN for ■Trademark Number■, and IPC for ■International Patent Classification■. More than half of the queries with the field search restrictions included application information related fields namely, AN, AP, and AD. One thing to note is that TL is used in conjunction with AB more than TL alone. It means that the users tend to look for the information either in the title or the abstract fields. Thus, it will be efficient in terms of the retrieval speed to index terms in the title and abstract fields together in addition to indexing them separately by the fields even it means some additional index storage space.

<Figure 3> Frequency Distribution of Field Search Restrictions in the Queries represented as the Percentage over all Queries with the Field Restrictions

3.7 Analysis of Multi-query Sessions

By analyzing the multi-query sessions, we found that the users had five intentions in reformulating queries such as paraphrasing, specialization, generalization, alternation, and interruption, which were also used by the web search engine users (Noriaki, et al., 2003).

<Table 10> Statistics concerning the Number of Queries, which Make-up All Sessions

Since there is no explicit session delimiter, we made an assumption that two or more consecutively issued queries by the same user will

Average number of queries in a session	1.04
Median number of queries in a session	1
Standard Deviation of number of queries per session	0.42
Maximum number of queries in a session	27
Minimum number of queries in a session	1

form one session. This assumption yielded 96,190 sessions from the query log. There were 2,202 multi-query sessions. The Table 10 shows the statistics about the sessions in terms of the number of queries. The sessions with only one query were 97.7% of all sessions.

The multi-query sessions were 2.3% of all queries. If we only consider the multi-query sessions, the average number of queries per session is 2.73, median is 2, standard deviation is 2.20, the minimum number of queries in a session is 2, and the maximum is 27. The Figure 4 shows the frequency distribution of the multi-query sessions excluding the single query sessions.

<Figure 4> Frequency Distribution of Number of Queries per Session Excluding the Single Query Sessions

We manually examined each multi-query session then identified the user intention in reformulating queries and also the types of query reformulation strategies that the users had used. We identified five user intentions and 17 query reformulation strategies.

<Table 11> Statistics concerning the User Intentions of Multi-Query Sessions

	Frequency	Percentage
Specialization	188	8.54
Generalization	384	17.44
Alternation	396	17.98
Paraphrasing	110	5.00
Interruption	529	24.02
Duplication	595	27.02

Out of 2,202 multi-query sessions, the queries in the respective 595 sessions were identical to each other. 595 sessions are 27.02% of all multi-query sessions. This means that the user had issued the same queries multiple times. The users might submit the duplicate queries to confirm the result from the original query. Whatever the reasons are, we believe that caching the initial query result during a single session will save the subsequent retrieval time. The Table 11 shows how the multi-query sessions were divided into five user intentions and the duplication category.

3.7.1 Specialization based Query Reformulation

According to Noriaki et al (2003), specialization was defined as the use of words to narrow the topic in a series of queries. We found two corresponding query reformulation strategies that the users of the

patent and trademark retrieval system adopted. The first strategy narrows the topic either by having additional query terms, which are conjunctively connected with the rest of the query or by extending a query term to include additional concept. An example of adding query terms with AND operator is as shown in the following. The Korean word ■플립' stands for 'flip' in English. '휴대폰' stands for 'cell phone'. '교체' stands for 'replacement'.

Original query: (플립<and>휴대폰)
Reformulated query: (플립<and>교체<and>휴대폰)

An example of extending a query term is usually done by replacing a single word or phrase query term into a phrase as shown in the following. The Korean word '부동산' stands for 'real estate' in English. '부동산사진' real estate photo'.

Original query: (부동산)
Reformulated query: (부동산사진)

The second strategy narrows the topic by adding field search restriction to the original query. An example of this query reformulation strategy is shown in the following. ■IPC■ is a field referred to as ■International Patent Classification■. Thus, the reformulated query specifies that the original query term should only occur in the IPC field.

Original query: (A47G23/06)
Reformulated query: ((A47G23)<in>(IPC))

The number of multi-query sessions, which belong to the first strategy, was 176. It is 93.62% of all multi-query sessions based on the specialization intention. There were 12 multi-query sessions, which belong to the second strategy. It is 6.38% of all multi-query sessions based on the specialization intention.

3.7.2 Generalization based Query Reformulation

Generalization was defined as the use of words to broaden the topic in a series of queries (Noriaki et al, 2003). We identified four corresponding query reformulation strategies in the query log. The first strategy broadens the topic by removing one or more query terms. An example of removing query terms is as shown in the following. An example of this strategy is as shown in the following. The Korean word '비타민' stands for 'vitamin' in English. '최진호' is a name of a person.

Original query: (비타민)<in>(TL,AB,CL)<and>((최진호)<in>(AP))
Reformulated query: ((최진호)<in>(AP))

The second strategy broadens the topic by removing field search restriction to the original query. An example of this query reformulation strategy is shown in the following. ■AP■ is a field referred to as ■Applicant Name or Code■. Thus, the reformulated query specifies that the original query term to occur in any field. The Korean word ■유재수' is a name of a person.

Original query: ((유재수)<in>(AP))
Reformulated query: (유재수)

The third strategy broadens the topic by adding ORed query term to the original query. An example of this query reformulation strategy is shown in the following.

Original query: (color<and>transient)
Reformulated query: ((color<or>colour)<and>(cti<or>transient))

The fourth strategy broadens the topic by decomposing a phrase or conjoined word into constituent words then joining them with AND operators. An example of this query reformulation strategy is shown in the following.

Original query: (CallerID)
Reformulated query: (Caller<and>ID)

The number of multi-query sessions, which belong to the first strategy, was 319. It is 83.07% of all multi-query sessions based on the generalization intention. There were 22 multi-query sessions, which belong to the second strategy. It is 5.73% of all multi-query sessions based on the generalization intention. There were 33 cases, which belong to the third strategy. It is 8.59% of all generalization cases. Finally, there were 10 cases, which belong to the fourth strategy. It is 2.60% of all generalization cases.

3.7.3 Alternation based Query Reformulation

Alternation was defined as the use of different words that does not change the general topic of the queries (Noriaki et al, 2003). Namely, the users■ intention is looking for the patents or the trademarks, with the similar query in comparison to the original query. We identified three query reformulation strategies in the query log, which belong to this intention. The first strategy seeks similar topic by replacing a part of the original query term. The Korean word ■파이프보호' stands for 'protecting pipes' in English. '파이프 마개' stands for 'pipe cover'.

Original query: (파이프 보호)<in>(TL,AB)
Reformulated query: (파이프 마개)<in>(TL,AB)

The second strategy changes the field search restriction in the original query. An example of this query reformulation strategy is shown in the following. 'AD' is a field referred to as 'Applicant Date'. Thus, the reformulated query looks for the same query term in different application date period. The Korean word '자판기' stands for 'automated vending machine'.

Original query: (자판기)<and>(AD>=20020713)<and>(AD<=20020720)
Reformulated query: (자판기)<and>(AD>=20020101)<and>(AD<=20021008)

The third strategy seeks similar topic by replacing all of the original query term. An example of this query reformulation strategy is shown in the following. The Korean word '퍼팅연습기' stands for 'putting exercise machine'. The Korean word '골프메트' stands for 'golf mat'.

Original query: (퍼팅연습기)
Reformulated query: (골프메트)

The number of multi-query sessions, which belong to the first strategy, was 341. It is 86.11% of all multi-query sessions based on the alternation intention. There were 9 multi-query sessions, which belong to the second strategy. It is 2.27% of all multi-query sessions based on the alternation intention. There were 46 cases, which belong to the third strategy. It is 11.62% of all alternation cases.

3.7.4 Paraphrasing based Query Reformulation

Paraphrasing was defined as the use of different words but keep the same meaning of the original queries (Noriaki et al, 2003). We identified four corresponding query reformulation strategies in the query log. Within the first strategy, the users replace either a part of the query or the whole query from one language to another. In this case, all occurrences of the query reformulation are from the Korean query terms to the English query terms. In the following example, the Korean word ■컴퓨팅' stands for 'computing' in English and thus there is no meaning change in the query reformulation.

Original query: (DNA<and>컴퓨팅)
Reformulated query: (DNA<and>computing)

In the second strategy, the cases of English query terms in the original are changed. The direction of case change can occur either from upper case to lower case and vice versa. This strategy does not involve Korean query terms as there is no case difference in Korean words. An example of this query reformulation strategy is shown in the following. The Korean word ■파라미터' stands for 'parameter'.

Original query: (qos<and>파라미터<in>AB)
Reformulated query: (QoS<and>파라미터<in>AB)

In the third strategy, the original query is replaced with the synonym in the reformulated query. An example of this query reformulation strategy is shown in the following. The Korean phrase '어류포획용어항' stands for 'device to capture fish'. The Korean word '어망' stands for 'fish net'.

Original query: (어류포획용어항)
Reformulated query: (어망)

In the fourth strategy, the query terms are swapped during the query reformulation process. An example of this query reformulation strategy is shown in the following. The Korean phrase '침대겸용책상' stands for 'a desk, which can be used as a bed'. The Korean phrase

‘책상겸용침대’ stands for ‘a bed, which can be used as a desk’.

Original query: (침대겸용책상)
Reformulated query: (책상겸용침대)

The number of multi-query sessions, which belong to the first strategy, was 59. It is 53.64% of all multi-query sessions based on the alternative intention. There were 8 multi-query sessions, which belong to the second strategy. It is 7.27% of all multi-query sessions based on the alternative intention. There were 10 cases, which belong to the third strategy. It is 9.09% of all alternative cases. There were 33 cases, which belong to the fourth strategy. It is 30.00% of all alternative cases.

3.7.5 Interruption based Query Reformulation

Interruption was defined as the use of different words that changes the topic of the original queries (Noriaki et al, 2003). We observed 529 cases belonging to this intention. We identified two query reformulation strategies in the query log, which belong to this intention. The first strategy looks for a totally different topic by replacing everything in the original query term. The Korean word ‘엔바이로테크’ is probably a name of a company or a product. Its phonetic English translation is ‘envirotech’. ‘강석주’ is a name of a person. Its phonetic translation is ‘Kang Seok Joo’.

Original query: ((엔바이로테크)<in>(AP))
Reformulated query: ((강석주)<in>(IN))

The second strategy partially replaces the topic of the original query. An example of this query reformulation strategy is shown in the following. The Korean word ‘사우나’ is ‘sauna’. ‘페인트’ is ‘paint’. ‘전기석’ is ‘electrically charged stone’. Thus, the reformulated query includes one common term and another replaced query term.

Original query: (사우나<and>페인트)
Reformulated query: (사우나<and>전기석)

The number of multi-query sessions, which belong to the first strategy, was 339. It is 64.08% of all multi-query sessions based on the interruption intention. There were 190 multi-query sessions, which belong to the second strategy. It is 35.92% of all multi-query sessions based on the interruption intention.

4 Conclusion and Further Research

By statistically analyzing the query log of the Korean patent and trademark retrieval system, we wanted to explore ways to improve the retrieval efficiency especially in terms of the retrieval speed. Unlike the web search and OPAC transaction logs that were used in the previous studies, the query log that we used lacked some of the typical information such as the corresponding results in response to

the submitted queries, the records of the user action after the retrieval results are returned, and any explicit clues indicating the beginnings and ends of the multi-query sessions. Thus, there were some limitations on what we can learn from the query log especially after the initial query is submitted. However, we made a reasonable assumption about the multi-query user sessions then went on to analyze the query reformulation activities..

Some of the retrieval efficiency improvement recommendations based on our systematic query log analysis are: 1) using the dedicated servers for top frequent users as nine percent of the users issued about one half of all queries; 2) directing the search requests to different servers depending on the number of query terms as the most of queries have three or less query terms and there were about one percent of the total queries, which had seven or more query terms; 3) optimizing the retrieval engine to speed-up particular field search as a limited number of specific field information was sought after about one half of the time, and 4) caching patent and trade retrieval results for the duplicate subsequent queries during the same multi-query sessions.

Caching the initial retrieval result will also improve the retrieval efficiency for the subsequent query if the subsequent query narrows the topic of the original query. This type of query reformulation is refereed to as specialization intention of the user. However, this will involve either automatically identifying a particular type of query intention or the user specifying his/her intention for reformulating the query to be submitted with the new query.

Our future work will focus on the analysis of the multi-query sessions. Firstly, we used a simple assumption of a session as a consecutive series of queries by the same user. However, this assumption cannot be accurate as the query log is the result of recording the submitted queries in a temporal sequence. Thus, there can be an intervening query by a user in the query log while another user is submitting a series of queries in a session. Thus, we will figure out more accurate way to identify the sessions to get the better picture of the users■ query reformulation strategies.

Secondly, our goal of semantically analyzing the sessions was mainly identifying the query reformulation strategy scheme and manually categorizing the multi-query sessions according to the scheme. In the future, we will explore ways to extract the particular semantic relations such as specialization and generalization amongst the query terms to investigate the possibility of generating a thesaurus from the query logs either for the users to refer when creating a query or for the system to utilize to improve its retrieval accuracy as well as coverage.

References

Blecic, D.D., Bagalore, N.S., Dorsch, J.L., Henderson, C.L., Koenig, M.H., & Weller, A.C. 1998. Using transaction log analysis to improve OPAC retrieval results. College and Research Libraries, 59(1), 39-50.

Hoelscher, C. 1998. How Internet experts search for information on the Web.

Proceedings of ebNet98 - World Conference of the World Wide Web, Internet, and Intranet, Charlottesville, VA, USA.

Jansen, B., Spink, A., & Sracevic, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36(2), 207-227.

Noriaki, K., Takeya, M., Miyoshi, H. 2003. Semantic Log Analysis Based on a User Query Behavior Model. Proceedings of the Third IEEE Conference on Data Mining, 107-115.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum, 33(1), 6-12.

Yoo, Jae Bok. 2004. Theories and Practices of Patent Information Retrieval. Seoul: Hyungsul Publisher.