2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Modeling Users' Information Goal Transitions and Satisfaction Judgment: Understanding The Full Search Process

Shandian Zhe*, Tian Xia* and Xueqi Cheng*
*Institute of Computing Technology
Chinese Academy of Science, Beijing, China 100080
Email: {zheshandian, xiatian}@software.ict.ac.cn cxq@ict.ac.cn

***Abstract*—To improve web search effectiveness and help personalized search applications, it is important to understand users' search process, especially the underlying information goal transitions and satisfaction judgment on result pages. Unlike previous work modeling the two types of hidden information separately, the paper proposes to simultaneously model them based on users' full search process, including both queries and clicks. Thus, a full model can be built up and the dependences between them can be leveraged. Specially, we employ a hierarchical conditional random field (HCRF) for learning and prediction, with fruitful search activity features proposed and leveraged. Experimental results show that our approach reaches a high overall precision (87%) and significantly outperforms the baseline methods. Moreover, our model is applied in a re-ranking application and shows that it can benefit personalized web search.**

***Keywords*-Search Process Model; Personalized Web Search; Query Topics; Click Model; HCRF**

I. INTRODUCTION

Due to the difficulty in describing the information needs properly by query and judging the relevance of web pages, users' search process is usually not a single-step activity. In a typical search scenario, users may need to reformulate queries consequently to better describe their information needs, and click multiple result pages before their information needs are satisfied. Therefore, users' search activity processes, including both issuing queries and clicking pages, actually reflect the hidden transitions between their information goals and hidden satisfaction on clicked pages. If we can capture this hidden information, we can better understand users' search process to further help web search applications. For example, one can improve the ranking of search results based on users' satisfaction on previous clicked results, or provide personalized web search by identifying users' information goal dynamically.

Previous work has proposed a series of click models [1][2][3] for users' search activities. These models treat a user's queries independently and focus on users' click actions within one query. Other works, like [4][5][6], are interested in detecting the transition between query topics during search. Although they leverage a lot of features from detailed queries, features about click actions and influences of a user's satisfaction are usually ignored or not fully utilized.

This paper considers modeling both information goal transitions and personal satisfaction judgment simultaneously for a user's full search process. The benefit is that the interactive influences between the two types of information can be fully leveraged and all useful features from search activities (e.g. issuing queries, clicking pages and various behavior patterns) can be considered. Specially, we first formalize the problem as predicting sequences of two types of hidden variables—*Information Goal Relation* and *Satisfaction Judgment* (Sec. II). A hierarchically structured conditional random field (HCRF) is then built to model the two sequences of hidden variables given the observed search process (Sec. III). A list of search activity features, which are far beyond previously published sets, are taken into account for classification (Sec. IV).

Experiments (Sec. V) show that our model derives a high overall precision (87%) and significant improvements over the baseline methods in precision. We also conduct extensive experiments to prove the effectiveness of our proposed new features. Moreover, we apply our model for re-ranking and the initial result shows that it can benefit customizing search results.

II. PROBLEM DEFINITION

We refer the transition between information goals hidden in queries to *Information Goal Relation*, with three values in its range:

- **The Same**: The two goals are identical.
- **Related**: Both of the goals belong to a more general goal (e.g. a user first searches for a product's quality and then for its price).
- **New**: The second goal is completely new compared to the first. (e.g. a user first searches for some movie and then for next day's weather.)

Similarly, we refer to a *Satisfaction Judgment* for each clicked page, with three values in its range:

- **Related & Not Enough**: The page is considered related but not enough to satisfy the user's information need.

This work is supported by the State Key Program of National Natural Science of China (Grant No. 60933005)

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.214

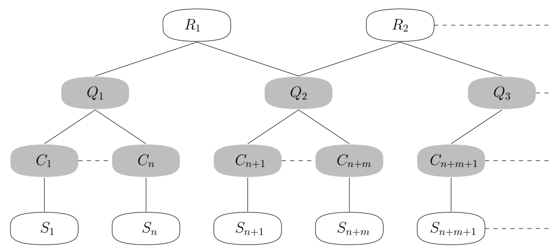

Figure 1. The concept hierarchy for modeling a user's search process (R_i denotes the *Information Goal Relation* between Q_i and Q_{i+1}, where Q_i represents i-th query; C_i represents i-th clicked page, with a *Satisfaction Judgment* S_i.)

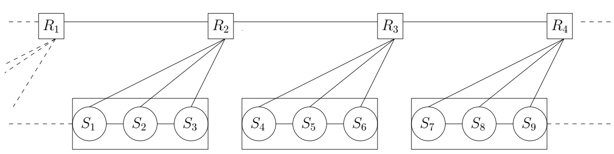

Figure 2. The HCRF's structure for modeling search process (R_i denotes i-th *Information Goal Relation*; S_i denotes i-th *Satisfaction Judgment* and ones inside the same hollowed rectangle belong to the same query. Observations are omitted for simplicity.)

- **Related & Enough**: The page is considered related and the user's information need is finally satisfied after viewing it.
- **Not Related**: The page is not related and no progress of the user's information need is made.

Thus, within a user's search activity sequence, for each pair of consecutive queries, there corresponds to an *Information Goal Relation*, and for each clicked page, there is a *Satisfaction Judgment*. Therefore, modeling search process is formalized as holistically predicting the variable values hidden in search activities, as shown in Figure 1.

III. Hierarchical Conditional Random Fields

In this section, a hierarchical conditional random field [7] (HCRF) is constructed for predicting *Information Goal Relation* and *Satisfaction Judgment*. The model is shown in Figure 2, with four types of cliques defined: (1) single node; (2) inter- *Information Goal Relation* edge; (3) inter-*Satisfaction Judgment* edge within one query; (4) *Information Goal Relation* and all *Satisfaction Judgment* within its first query. Thus, a list of search activity features can be defined over all these cliques, as shown in Section IV.

IV. Features for HCRFs

A. Notations for Feature Definition

We denote a serial of *Information Goal Relation* as $\{r_1, r_2, \ldots\}$ with value range R and *Satisfaction Judgment* as $\{s_1, s_2, \ldots\}$ with value range S. For each r_i ($i \geq 1$), the associative i-th and $i+1$-th query are denoted as q_i and q_{i+1}; For each $s_i (i \geq 1)$, the associative snippet and page are denoted as n_i and p_i, respectively.

Note that for each feature, a group of feature functions are generated by multiplying an indicator function. For example, if a feature $f(x)$ is chosen for predicting y, the feature functions are derived as $f(x, y) = f(x) \cdot I_v(y)$, where $I_v(y)$ is the indicator function which equals 1 if $y = v$ and 0 otherwise and v iterates over y's value range.

B. Temporal Features

The features are about intervals of a user's search activities. Previous published features of this type [5] are mainly about issuing queries, however, ones about click actions are also explored here:

- inter-query time in minutes for r_i , also with a group of thresholds
- inter-click time in seconds within one query for s_i, also with a group of thresholds
- time-difference in minutes between click and next query, also with a group of thresholds. They are defined for r and s within r's first query.
- ratio of two consecutive inter-click times within one query for s_i and s_{i+1}: it reflects a user's preference differences toward two clicked pages.

C. Word and Character Features

While previous work only concern query similarities [5][6], we consider both query and snippet similarities:

Features based on inter-queries (q_i, q_{i+1}) for r_i

- normalized edit distance, also with some thresholds
- common character number starting from left (right)
- common word number starting from left (right)
- common word number
- Jaccard distance on word sets

Features based on q_j and n_i's title for s_i in q_j: the same as the above list

Features based on q_j and n_i's content for s_i in q_j

- common word number
- Jaccard distance on word sets
- cosine distance on word vectors, also with some thresholds

Features based on q_j and n_i for s_i in q_j

- Jaccard distance on word sets
- cosine distance on word vectors, also with some thresholds

Features based on n_i and top-10 snippets of next query q_j for s_i and r_j

- cosine distance on title (content, snippet) word vectors, also with a group of thresholds

Features based on (n_i, n_{i+1}) for s_i and s_{i+1}

- normalized edit distance on titles, also with a group of thresholds
- titles' common character number starting from left (right)
- titles' common word number starting from left (right)

- common word number on titles, contents and snippets respectively
- Jaccard distance of word sets on titles, contents and snippets respectively
- cosine distance of word vectors on contents and snippets respectively, also with some thresholds

Note that snippet contents are mixed and much longer than queries and snippet titles thus it may not be appropriate to measure character similarities between them. Thereby, words similarities are only leveraged here.

D. Behavior Pattern Features

A group of users' behavior patterns during search are summarized as follows:

- the rank of the last clicked page in q_i for r_i
- number of times clicking "next" link in q_i's search result page for r_i, also with some thresholds
- number of clicked pages in q_i for r_i, also with some thresholds
- the rank of p_i for s_i, also with some thresholds
- whether p_i is the last clicked page in the query for s_i
- rank difference of (p_i, p_{i+1}) for s_i and s_{i+1}, also with a group of thresholds

E. Connection Features

This feature type describes the dependences among variables. Some formerly defined features (like ones for both s and r) can also be assigned to this type.

- $I_{u,v}(s,r)$, where $u \in S, v \in R$ and $I_{u,v}(x,y)$ is the binary indicator function. s associates a clicked page within the first query of r.
- $I_{u,v}(s,r)$, where $u \in S, v \in R$. s associates the last clicked page within r's first query.
- ratio of s equal to *Not Related* within the first query of r. The feature is defined for r and all s within r's first query. Intuitively, the ratio approaches 1 means that the user does not satisfy the returned search results thereby may reformulate the query and tries again.
- $I_{u,v}(r_i, r_{i+1})$, where $u, v \in R$
- $I_{u,v}(s_i, s_{i+1})$, where $u, v \in S$
- $I_u(r)$, where $u \in R$
- $I_u(s)$, where $u \in S$

V. Experiment

A. Dataset

To obtain an effective dataset, we first developed a Firefox plug-in for recording and labeling users' search activities and then invited six volunteers for data collection: they used the search engine (Google) as normal, but when issuing a query or reading a clicked page, they were asked to label *Information Goal Relation* or *Satisfaction Judgment*. The volunteers could delete any records relating to their privacies. The whole process lasted for four months and a labeled dataset containing 7134 queries and 7032 clicked pages was finally collected. The dataset contains 6233 distinct queries, involving various information goals such as traveling, food, music, tickets and computer technology. Table I shows the distribution of the hidden variables.

Table I
Distribution of hidden variables

Variable Type	Distribution	
Information Goal Relation	The Same	33%
	Related	23%
	New	44%
Satisfaction Judgment	Related & Not Enough	31%
	Related & Enough	45%
	Not Related	24%

Table II
Precision of Classification Result

Methods	baseline1(%)	baseline2(%)	all(%)
The Same	56	76	**89**
Related	59	63	**82**
New	80	81	**87**
Related& Not Enough	64	62	**80**
Related & Enough	68	69	**88**
Not Related	54	54	**75**

B. Methodology and Classification Result

The leave-one-out cross-validation is adopted for evaluation, that is, every five persons' data are used for learning and the left one's is for evaluation. The average of all six evaluations is regarded as the final result.

Two baselines are chosen for comparison: baseline1 only contains thresholds of temporal features: inter-query and inter-click time thresholds; baseline2 contains the best features published for detecting topic change boundaries [5], mainly including inter-query features in Section IV. Since the paper does not explore usage of extra knowledge, features from web page content and huge query logs in [5][6] are not incorporated. Similar to baseline1, baseline2 also utilizes inter-click time thresholds for *Satisfaction Judgment*.

Our model utilizes all features in Section IV. Thus totally three feature sets are incorporated into HCRFs separately for comparison. The result is shown in Table II and all differences are statistically significant.

It can be seen that our model combining all features in Section IV significantly outperforms baseline1 and baseline2 in precision. For overall precision of classifying the two hidden variables, our model reaches 87% and derives a 34% and 18% improvement over baseline1 (65%) and baseline2 (74%) respectively.

C. Contribution of Feature Types

The contribution of each feature type is examined separately and listed in Table III, where 'Sa', 'Re', 'Ne' are short for *The Same*, *Related* and *New*, 'NEn', 'En', 'NR' are short for *Related & Not Enough*, *Related & Enough* and *Not Related*. All differences are statistically significant.

Table III
PRECISION OF EACH FEATURE TYPE(INFORMATION GOAL RELATION/SATISFACTION JUDGMENT)

Feature Type	Sa/NEn(%)	Re/En(%)	Ne/NR(%)
baseline1	56/64	59/68	80/54
baseline2	76/62	63/69	81/54
temporal	58/65	62/68	85/57
word & character	82/72	70/76	84/69
behavior pattern	67/56	42/74	55/52
connection	53/56	57/69	54/43
without connection	86/75	78/84	85/71

It can be seen that only using temporal features is slightly better than baseline1 and this may be due to more features and the learned thresholds which are better than pre-assigned thresholds.

Word and character features significantly outperform baseline2 thereby it demonstrates that the extending features of this type (such as ones about snippets) are effective for improving performance.

Behavior patterns and connection features also prove their effects for classification. However, just using either type leads to a poor result: behavior pattern features perform badly for *Related*; connection features cannot recognize *Not Related* well. This may because both types are not subtle enough to capture the classification boundaries, especially for *Related* and *Not Related*. Finally, a model combining all features except the connection type in Section IV is built and tested in the last row of Table III. The model is equivalent to two independent classifiers for *Information Goal Relation* and *Satisfaction Judgment* respectively since no dependences among variables are considered. The statistical significance test shows that although with a quite high precision, the model is still inferior to our model using all features in Table II. Thus, the advance of modeling the two types of variables together is proved, since the dependences between them do help improve the classification precision.

D. Re-Ranking Application

Given page p, unrelated page set S_1 and related page set S_2, our re-ranking function is defined as:

$$\begin{aligned} rank(p) \quad = \quad & (1-\alpha-\beta)\cdot\log(ori(p)) \\ & -\alpha\cdot KL(p,S_1)+\beta\cdot KL(p,S_2) \end{aligned} \tag{1}$$

where $0 \leq \alpha, \beta \leq 1$, $ori(p)$ is p's original rank and KL is short for Kulback-Leibler divergence. We re-rank the top 20 results with this formula when the *Information Goal Relation* is *The Same*. We take S_1 as the clicked pages classified as *Not Related* in the last query and S_2 as ones classified as *Related & (Not) Enough*.

25 different queries are randomly selected from the dataset for re-ranking. For each query, the volunteer issuing it is asked to label the relevance of the top 20 returned pages. The MAP of both the original and the new are calculated for comparison, as shown in Figure 3. We can see most of the new rank results become better thereby our model is proved to benefit customizing search results.

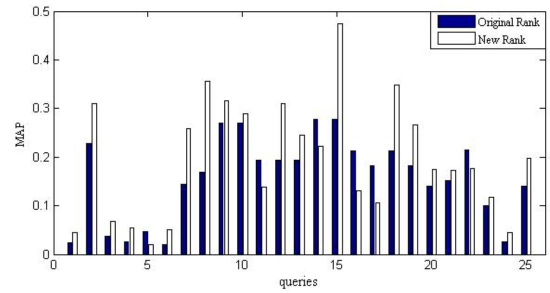

Figure 3. MAP of old and new results

VI. FUTURE WORK

In the future work, we would like to explore extra knowledge such as web page contents and huge query logs to further improve our model's performance.

REFERENCES

[1] N. Craswell, O. Zoeter, M. Taylor and B. Ramsey, *An Experimental Comparison of Click Position-Bias Models*, *In Proc. of WSDM'08*, California USA: ACM Press, Februray 2008, pp. 87-94.

[2] F. Guo, C. Liu, and Y. Wang, *Efficient Multiple-Click Models in Web Search* , *In Proc. of WSDM'09*, Barcelona Spain: ACM Press, February 2009, pp. 124-131.

[3] F. Guo, L. Li, and C. Faloutsos, *Tailoring Click Models to User Goals*, *In WSCD workshop of WSDM'09*, Barcelona Spain: ACM Press, February 2009, pp. 88-92.

[4] T. Lau and E. Horvitz, *Patterns of Search: Analyzing and Modeling Web Query Refinement*, *In Proc. of UM'99*, Banff: Springer Press, June 1999, pp. 119-128.

[5] R. Jones, K.L. Klinkner, *Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs*, *In Proc. of CIKM'08*, California USA: ACM Press, October 2008, pp. 699-708.

[6] P. Boldi, F. Bonchi, C. Castillo and S. Vigna, *From 'Dango' to 'Japanese Cakes': Query Reformulation Models and Patterns*, *In Proc. of WIC'09*, Milan Italy: IEEE Computer Society, September 2009, pp.183-190.

[7] J. Lafferty, A. McCallum and F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, *In Proc of ICML '01*, Williamstown MA: ACM Press, June 2001, pp. 282-289.