

Acquiring Explicit User Goals From Search Query Logs

Markus Strohmaier

Graz University of Technology
and Know-Center

markus.strohmaier@tugraz.at

Peter Prettenhofer

Graz University of Technology

peter.prettenhofer@
student.tugraz.at

Mark Kröll

Graz University of Technology

mkroell@tugraz.at

Abstract

Knowledge about user goals is crucial for realizing the vision of intelligent agents acting upon user intent on the web. In a departure from existing approaches, this paper proposes a novel approach to the problem of user goal acquisition: The utilization of search query logs for this task. The paper makes the following contributions: (a) it presents an automatic method for the acquisition of user goals from search query logs with useful precision/recall scores (b) it provides insights into the nature and some characteristics of these goals and (c) it shows that the goals acquired from query logs exhibit traits of a long tail distribution.

1. Introduction

To realize the vision of intelligent, goal-oriented agents on the web, agents must have programmatic access to the set and variety of human goals, in order to reason about them and to provide services that help satisfy users' needs. In Berner's Lee vision, an agent aiming to "plan a trip to Vienna" would need to have some means to understand that "plan a trip" is likely to involve a set of other goals or services, such as "contact a travel agency" and "book a hotel". This type of knowledge has been characterized as commonsense knowledge, i.e. knowledge that humans are generally assumed to possess, but which is extremely difficult for computers to acquire. Examples of current research projects aiming to capture and organize commonsense knowledge, including knowledge about human goals, are CyC [6] or Openmind / ConceptNet [7]. However, existing attempts suffer from two main problems: 1) the *goal acquisition problem (or bottleneck)*, which refers to the costs associated with knowledge acquisition and 2) the *goal coverage problem*, which refers to the difficulty of capturing the tremendous variety and range in the set of human goals. These problems have hindered progress in capturing broad knowledge about human goals, and have hindered the

development of intelligent agents, services and applications on the web.

On the web, search engines represent a primary instrument through which users exercise their intent today. This allows search queries to indirectly convey knowledge about users' goals and intentions, which are usually latent, implicit, dynamic and private. Given that existing attempts to capture knowledge about human goals are usually limited, an interesting question in the context of search is: *Can we automatically acquire knowledge about a large variety of user goals from search query logs?* In this paper, we study *if, how and to what extent* it is possible to automatically acquire knowledge about explicit user goals (such as "book a hotel") from search query logs.

2. Human Subject Study

In order to gauge the results of an automatic acquisition approach addressing this problem, we first conducted a human subject study aiming to 1) define the notion of explicit user goals more rigorously and 2) to learn about its principal agreeability.

Definition of Explicit User Goals: We define queries containing explicit user goals in the following way:

A search query is regarded to contain an explicit user goal (or short: explicit goal) whenever the query 1) contains at least one verb and 2) describes a plausible state of affairs that the user may want to achieve or avoid in 3) a recognizable way.

An example of such a query would be "book a hotel". A query does not contain an explicit goal when it is difficult or extremely hard to elicit some specific goal from the query. Examples include blank queries, or queries such as "car" or "travel", which embody user goals on a very general, ambiguous and mostly implicit level.

Questionnaire Design: To explore the utility of this definition, we have conducted a questionnaire in which four human subjects (Computer Science

graduate students) were instructed to manually label 3000 queries randomly obtained from the AOL search query log [8] (after a number of sanitization and pre-processing steps were performed). The subjects were required to independently answer a single question for each of the 3000 queries. The question for each query followed this schema: *Given a query X, Do you think that Y (with Y being the first verb in X, plus the remainder of X) is a plausible goal of a searcher who is performing the query X?* To give two examples:

Given query: “how to increase virtual memory”
Question: Do you think that “increase virtual memory” is a plausible goal of a searcher who is performing the query “how to increase virtual memory”? Potential Answer: Yes
Given query: “boys kissing girls” Question: Do you think that “kissing girls” is a plausible goal of a searcher who is performing the query “boys kissing girls”? Potential Answer: No

After the question-answering task, we assigned the answers for each query to the corresponding categories ourselves in the following way: each answer “Yes” resulted in classifying the query as a “*query containing an explicit goal*”, each answer “No” resulted in classifying the query as a “*query not containing an explicit goal*”. The results are reported next.

Agreeability of Constructs: We calculated a function $e(q)$ per query, which is the percentage of human subjects who labeled a given query as containing an explicit goal (cf. [5]). The chart in Figure 1 shows that 243 queries out of 3000 have been labeled as containing an explicit goal by all 4 subjects (8.1%, right most bar), and 134 queries as containing an explicit goal by 3 out of 4 subjects. The majority of queries (79.2%, left most bar) has been labeled as not containing an explicit goal unanimously by all subjects. A relatively small number of queries was controversial (middle bar, 3.3%). Figure 1 shows that $e(q)$ approximates a dichotomous agreement distribution, which provides preliminary evidence for the agreeability of our constructs. To further explore agreeability, we calculated the inter-rater agreement κ [2] between all pairs of human subjects A, B, C and D. Cohen’s κ measures the average pairwise agreement corrected for chance agreement when classifying N items into C mutually exclusive categories. The κ values in our human subject study range from 0.65 to 0.76 (see Figure 1). Both measures combined, the inter-rater agreement κ and the distribution of $e(q)$, hint towards a principle (yet not optimal) agreeability of our construct definition. In the remainder of this paper, we use these results to inform the development of an automatic classification approach.

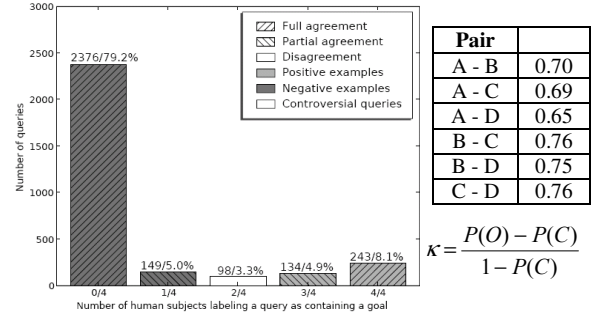


Figure 1. $e(q)$ Distribution and

3. Acquiring Explicit User Goals

Based on the human subject study, we now introduce an inductive classification approach that aims to perform the task of classifying queries into one of the two categories (containing/not containing an explicit goal) automatically.

Training Set: We have created a manually labeled dataset for the purpose of training an automatic classification approach. The manually labeled dataset is based on the majority vote among the human subjects of the human subject study presented previously. Out of the 3000 labeled queries, the negative examples were defined by the two bars on the left hand side of Figure 2 (2525 total), and the positive examples were defined by the two bars on the right hand side (377). The bar in the middle represents the controversial queries which were removed.

The approach for classifying queries consists of two basic steps: POS tagging and classification.

POS Tagging: We can assume that queries containing explicit goals can, to some extent, be identified by the occurrence of certain syntactical, part-of-speech patterns. To investigate this, we used a Maximum Entropy Tagger for part-of-speech tagging all queries. We used the Penn Treebank tag set containing 36 word classes which provides a simple yet adequately rich set of tag classes for our purpose.

Feature Set Description: The following feature types were utilized:

- **Part-of-Speech Trigrams:** Each query can be translated from a sequence of tokens into a sequence of POS tags. Trigrams were generated by moving a fixed sized window of length 3 over the POS sequence. The sequence boundaries were expanded by introducing a single marker (\$) at the beginning and at the end allowing for length two POS features. The query “buying/VBG a/DT car/NN” would yield the following trigrams:

\$ VBG DT; VBG DT NN; DT NN \$

- **Stemmed unigrams:** Queries can be represented as binary word vectors or ‘Set of Words’ (SoW). The Porter stemming algorithm was used for word conflation and stopwords were removed.

We chose a linear Support Vector Machine (SVM) using all the features as our classification method. The performance of this approach is discussed next.

Evaluation: Table 1 presents the confusion matrix on the manually labeled dataset and corresponding True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) scores.

Table 1. Confusion matrix

Classified as →	Containing an Explicit Goal	Not Containing an Explicit Goal
Containing an Explicit Goal	239 (TP)	138 (FN)
Not Containing an Explicit Goal	73 (FP)	2452 (TN)

Our approach achieves a precision of 0.77, a recall of 0.63 and an F1 score of 0.69. All values refer to the class that represents queries containing goals. A precision of 77% means that in 77% of cases, our approach agrees with the majority of human subjects. These results represent a significant improvement over previous approaches [9]. Although there are slight differences in the evaluation procedure and the type of knowledge captured, the precision of explicit goals acquired with our approach is roughly comparable to precision scores reported for the ConceptNet commonsense knowledge database.

4. Results

In the following, we present the results of applying our automatic classification approach to a pre-processed version of the entire AOL search query log containing more than 20 million search queries.

Selected Statistics: Applying our automatic classification method yielded a result set containing explicit user goals consisting of 118.420 queries, 97.454 of them unique. With a precision of 77%, the result set comprises an estimated 75.039 *true positives* (*actual queries containing explicit goals*).

The 20 most frequent queries from the result set are presented in Table 2. Each example is accompanied by the rank and the number of different users who submitted the query (frequency). Queries containing the token “http” were filtered out and those queries containing expletives / objectionable content were replaced by “deleted”. Some of the most frequent queries containing goals relate to commonsense knowledge goals, such as “lose weight”, “get pregnant” or “listen to music”, which provides some evidence of the

suitability of search query logs for the knowledge acquisition task. Yet, the bias introduced by the corpus (search queries) and the population (i.e. AOL users) deserves attention: Many frequent queries deal with web-related or AOL specific issues, such as the queries “add screen name” or “cancel aol service”. Entries such as “wedding cake toppers”, “pimp my ride”, and “skating with celebrities” represent false positives.

Table 2. 20 most frequent goals

Nr.	Query	#Users	Nr.	Query	#Users
1	add screen name	205	11	cancel aol service	54
2	create screen name	137	12	pimp my myspace	53
3	rent to own	120	13	cancel aol account	50
4	listen to music	108	14	“deleted”	49
5	pimp my space	102	15	”deleted”	48
6	pimp my ride	97	16	how to lose weight	47
7	assist to sell	93	17	how to get pregnant	47
8	wedding cake toppers	64	18	change my password	46
9	skating with celebrities	58	19	discover credit card	46
10	lose weight fast	56	20	check my computer	43

If search query logs would be utilized for knowledge acquisition, a relevant question to ask is: *How diverse is the set of goals contained in search query logs?* The diversity of goals would ultimately constrain the utility of a given dataset for expanding existing knowledge bases. In order to explore this question, we present a rank/frequency plot of the data depicted in Table 2. In Figure 2, goals are plotted according to their rank and the set of different users who share them.

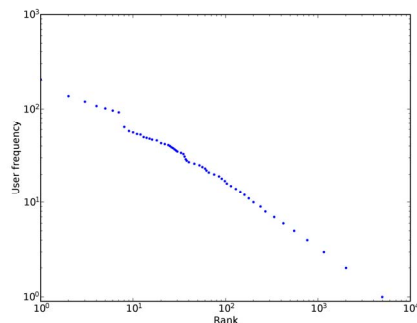


Figure 2. Rank-frequency plot of goals

The distribution in Figure 2 shows that while there are very few popular goals, a majority of goals is shared by only a few users. In other words, the curve approximates a power-law distribution, implying the existence of a long tail effect of user goals. This suggests that the explicit goals in the result set are diverse and cover a broad range of different goals.

Qualitative Analysis: We selected an arbitrary set of verbs and corresponding goals for more detailed inspections. In Table 3, the 10 most frequent goals which contain the verbs “get”, “make”, “change” or “be” are listed. Frequency refers to the occurrence of the goal in the result set.

The goals in Table 3 are the result of identifying the first verb in a query containing a goal, and truncating any tokens prior to this verb. Queries marked with a “*” represent queries that are contained in ConceptNet’s commonsense knowledge base (v2.1) as well. Many goals in Table 3 are related to existing commonsense knowledge goals, such as “be pregnant”, “be rich” or “be funny”.

Table 3. 10 most frequent goals containing “get”, “make”, “change” and “be”

#	Verb: get	Verb: make	Verb: change	Verb: be
1	get pregnant* (141)	make money* (87)	change my password (100)	be anorexic* (26)
2	get rid of ants (28)	make your own website (43)	change my screen name (38)	be pregnant* (19)
3	get out of debt planner (19)	make money at home (41)	change screen name (32)	be bulimic (12)
4	get rich or die tryin (17)	make money fast (39)	change my aol password (28)	be rich* (11)
5	get rid of love handles(17)	make money online (34)	change password (24)	be emo (8)
6	get married (15)	make the band 3 (30)	change my profile (21)	be funny* (8)
7	get rich* (15)	make money from home (25)	change your name (21)	be happy* (8)
8	get rich with trump (15)	make new screen name (24)	change* (20)	be sexy* (7)
9	get out of debt* (15)	make up (23)	change my email address (17)	be in love* (7)
10	get rid of moles (14)	make out (21)	change aol password (14)	be an actress (7)

5. Related Work and Conclusions

In previous research, He et al. [3] have studied the acquisition of explicit user goals from search *result snippets* (i.e. the segments of text listed on the result pages of search engines). Our work is different in the sense that it studies search queries themselves as a source of explicit goals, which can be suspected to better reflect user intent. Broder’s high level taxonomy of search intent, proposing a distinction between three classes of search goals, has stimulated a series of follow-up research on category refinement and automatic query categorization [4][5]. While previous research has achieved considerable progress in the *categorization of queries* into high-level goal

taxonomies serving a primarily *functional purpose* (to improve search), this work focuses on the *acquisition of goal instances (explicit goals)* from search query logs for *knowledge capture* purposes.

Our work shows that search query logs have the potential to address the two problems (goal acquisition and goal coverage) of acquiring knowledge about human goals on the web. In a departure from existing approaches, we present an automatic classification approach and experimental results that introduce search query logs as a *feasible*, yet largely *untapped* resource for this task.

Acknowledgements: This work is funded by the FWF Austrian Science Fund Grant P20269 TransAgere and the Know-Center. The Know-Center is funded within the Austrian COMET Program.

REFERENCES

- [1] A. Broder, A taxonomy of web search, SIGIR Forum, vol. 36, no. 2, pp. 3-10, 2002.
- [2] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, (20)1:37, 1960.
- [3] K.Y. He, Y.S. Chang and W.H. Lu. Improving identification of latent user goals through search-result snippet classification. In Proceedings of the International Conference on Web Intelligence, 683-686, IEEE Computer Society, 2007.
- [4] B.J. Jansen, D.L. Booth and A. Spink. Determining the informational, navigational, and transactional intent of web queries. Information Processing and Management, (44)3:1251-1266, Elsevier, 2008.
- [5] U. Lee, Z. Liu and J. Cho. Automatic identification of user goals in web search. In Proceedings of WWW 2005, ACM Press, New York, USA, 2005.
- [6] D.B. Lenat. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, (38)11:33-38, 1995.
- [7] H. Liu and P. Singh. ConceptNet - A practical commonsense reasoning tool-kit. BT Technology Journal, (22)4:211-226, 2004.
- [8] G. Pass, A. Chowdhury and C. Torgeson. A picture of search. In Proceedings of the 1st International Conference on Scalable Information Systems, ACM Press New York, NY, USA, 2006.
- [9] M. Strohmaier, P. Prethenhofer and M. Lux. Different degrees of explicitness in intentional artifacts - studying user goals in a large search query log. In Proceedings of the CSKGOI’08 Workshop on Commonsense Knowledge and Goal Oriented Interfaces, held in conjunction with IUI’08, Canary Islands, Spain, 2008.