Acquiring Knowledge About Explicit User Goals from Search Query Logs

Markus Strohmaier
Graz University of Technology and
Know-Center
Inffeldgasse 21a
8010 Graz, Austria
markus.strohmaier@tugraz.at

Peter Prettenhofer
Know-Center
Inffeldgasse 21a
8010 Graz, Austria
pprett@know-center.at

Mark Kröll
Graz University of Technology
Inffeldgasse 21a
8010 Graz, Austria
mkroell@tugraz.at

ABSTRACT
Access to knowledge about user goals represents a critical component for realizing the vision of intelligent agents acting upon user intent on the web. Yet, the acquisition of knowledge about user goals represents a major challenge. In a departure from existing approaches, this paper proposes a novel perspective for knowledge acquisition: The utilization of search query logs for this task. The research presented in this 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, thereby providing access to a broad range of user goals. We conclude that our work has implications for open research problems such as common sense knowledge acquisition and studies of user intent in query logs.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms
Algorithms, Experimentation

Keywords
Knowledge acquisition, explicit user goals, query log analysis

1. INTRODUCTION

1.1 Motivation
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 [18][27]. In Berner's Lee vision, an agent aiming to, for example, "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 [20]. Examples of current research projects aiming to capture and organize commonsense knowledge, including knowledge about human goals, are CyC [15] or Openmind / ConceptNet [28], which utilize human knowledge engineering [15], volunteer-based [20], game-based [18] or semi-automatic approaches [7] for knowledge acquisition. However, existing attempts suffer from two main problems: 1) the *goal acquisition problem (or bottleneck)*, which refers to the costs associated with knowledge acquisition [18] and 2) the *goal coverage problem*, which refers to the difficulty of capturing the tremendous variety and range in the set of human goals [7]. 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 a large number of users exercise their intent today. During search, users frequently refine, generalize and evolve their daily informational needs by formulating and executing a sequence of search queries, thereby leaving "traces of intent" [30]. 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 expensive and limited in scope, an interesting question in the context of search is: *Can we automatically acquire knowledge about a large variety of user goals from search queries, a natural byproduct of human activity on the web?*

While search query logs have been utilized successfully for knowledge acquisition in a range of different contexts [22], they

have not been used to acquire knowledge about explicit user goals, partly because query logs pose a number of challenges: The majority of search queries is usually short [22] and ambiguous [1], they convey user goals at different degrees of intentional explicitness [31], and they are often consisting of arbitrary concatenations of terms that frequently contain misspellings. Yet, recent research suggests that a number of search queries actually contains explicit statements of goals [31], and that the space of queries in search query logs is vast and topically diverse [23]. This would make query logs a particularly useful resource for acquiring knowledge about diverse user goals.

For illustration, Table 1 gives some examples of actual queries containing/not containing explicit statements of goals (or "*explicit goals*" from here on) obtained from a real world query log [23]

Queries containing explicit statements of goals	Queries not containing explicit statements of goals
"sell my car"	"Mazda dealership"
"play online poker"	"online games"
"find home to rent in Florida"	"Miami beach houses"
"passing a drug test"	"drug test"
"raising your credit score"	"credit cards"

Table 1 Examples of queries obtained from [23]

If it were possible to (1) distinguish queries containing explicit goals from those that don't and (2) identify or infer associations between user goals, it would be possible to construct or extend *broad* commonsense knowledge bases from an abundant, renewable resource (i.e. search query logs) in a cost-effective way, thereby potentially addressing the two problems identified earlier.

This paper focuses on task (1), that is distinguishing queries containing explicit goals from those that do not. We study *if*, *how* and *to what extent* it is possible to automatically acquire knowledge about explicit user goals from search query logs. Tapping into search query logs as a resource for knowledge acquisition about user goals could help lower costs often associated with knowledge acquisition, and also increase the coverage of existing approaches (such as CyC and ConceptNet). Due to the novelty and complexity of task (1), addressing task (2) is beyond the scope of this paper.

1.2 Contributions

This paper suggests that search query logs (sometimes also characterized as "databases of intentions"[1]) represent a *viable*, yet largely *untapped* resource for acquiring knowledge about explicit user goals.

Specifically, this paper makes the following contributions: In Section 2, we present the results of a human subject study that aimed to develop a useful definition for "queries containing explicit user goals". Based on this definition, we present an automatic classification approach that is capable of acquiring queries containing explicit goals with reasonable precision/recall scores in Section 3. In Section 4 and 5, we present results from experiments applying and evaluating the automatic classification approach resulting in the acquisition of an estimated set of 75.000 explicit user goals from a large search query log. Section 6, and 7 discuss related work and threats to validity. Section 8 concludes with potential implications of our research and future work. The overall contribution of the research presented in this paper is the presentation of a method for acquiring knowledge about user goals from search query logs, the discussion of the nature and some characteristics of the goals contained in search query logs and the empirical corroboration that goals in search query logs exhibit huge variety, and are approximating a power law ("long tail") distribution.

[1] http://battellemedia.com/archives/000063.php, last accessed on May 30, 2008

2. RESULTS OF HUMAN SUBJECT STUDY

We approach the problem of acquiring knowledge about user goals from search query logs as a binary classification problem, aiming to separate queries containing explicit user goals from other queries. This classification problem has been shown to represent an orthogonal problem to existing approaches to search intent categorization [31]. In order to gauge the results of an automatic classification approach addressing this problem, we conducted a human subject study aiming to 1) define our constructs more rigorously and 2) to learn about their principal agreeability.

2.1 Explicit User Goals

Based on work that emphasizes the crucial role of verbs in explicit goal formulations ([17],[26]), 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 (cf. [26]) in 3) a recognizable way [30].

"Recognizable" refers to what [13] defines as "trivial to identify" by a subject within a given attention span. "Plausible" refers to an external observer's assessment whether the goal contained in a query could likely represent the goal of a user who formulates the given query. It is important to note that it would be rather difficult to completely verify this assessment solely based on data from an anonymous query log due to the inherent *goal verification problem* of such a task [30]. However, the objectives of this work are more modest: In this paper, we are interested in acquiring *plausible* user goals for *knowledge capture* purposes. An advantage of acquiring broad knowledge about *plausible* user goals is that it can put constraints on the space of *possible* user goals, which plays a role in, for example, goal recognition [10] or query disambiguation [1].

"Queries containing explicit goals" according to our definition can be related to what other researchers have characterized as "better queries", or queries that have "more precise goals" (R. Baeza-Yates at the "Future of Web Search" Workshop 2006, Barcelona). 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.

2.2 Questionnaire Design

To explore the agreeability of our definition, we have conducted a pilot questionnaire in which 4 human subjects (Computer Science graduate students) were instructed to manually label 100 queries each. For each query, the subjects were supposed to decide whether the query contains an explicit goal or not. The theoretic nature of our distinction has caused several difficulties in the labeling task (uncertain classification, lack of understanding the definition, disagreements), which were addressed in the following way: In a subsequent, more comprehensive human subject study, we have replaced the labeling task with a question-answering task. In this task, the subjects were required to independently answer a single question for each of 3000 queries randomly obtained from a real world search query log after sanitization and pre-processing steps were performed (more details on our dataset are presented in Section 4.1). 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:

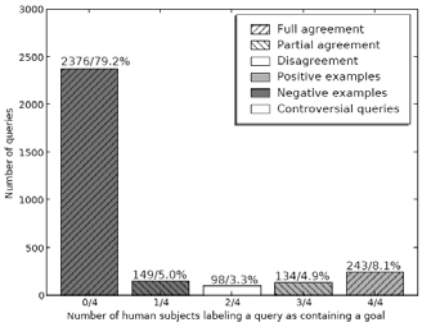
> Given query: "how to increase virtual memory" Question: Do you think that "increase virtual memory" is a plausible goal of a searcher who is performing the query "how to increase virtual memory"? Potential Answer: Yes
>
> Given query: "boys kissing girls" Question: Do you think that "kissing girls"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

These questions were designed to learn about the specific query formulations that human subjects associate with goals (such as "how to …"). 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 of this human subject study are reported next.

2.3 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. [14]). 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 have been labeled as containing an explicit goal by 3 out of 4 subjects. This corroborates the assumption that query logs contain a small number of queries that contain explicit goals. The majority of queries (79.2%, left most bar) has been labeled as not containing an explicit goal unanimously by all subjects. This shows that a relatively small number of queries was controversial (middle bar, 3.3%). 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 [5] 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. In the formula depicted in Figure 1, P(O) is the proportion of times that a hypothesis agrees with a standard (or another labeler), and P(C) is the proportion of times that a hypothesis and a standard would be expected to agree by chance. The κ value is constrained to the interval [-1,1]. A κ value of 1 indicates total agreement, 0 indicates agreement by chance and -1 indicates total disagreement. 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 definition.



Pair	κ
A - B	0.70
A - C	0.69
A - D	0.65
B - C	0.76
B - D	0.75
C - D	0.76

$$\kappa = \frac{P(O) - P(C)}{1 - P(C)}$$

Figure 1 e(q) Distribution and Interrater-Agreement κ

In the remainder of this paper, we use these results to inform the development of an automatic classification approach.

3. ACQUIRING EXPLICIT USER GOALS

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 to aid knowledge acquisition. Current approaches to search query classification such as topic and search intent classification use a variety of features for their tasks. Lee et al. [14] for example use click-through and anchor text information to categorize queries as either navigational or informational. Gravano et al [9] categorize queries according to geographical locality by utilizing a search engine to overcome the inherent sparsity problem when dealing with (short) queries. For acquiring knowledge about explicit goals contained in search query logs, it is necessary to examine linguistic features of the queries themselves.

To study if and how queries containing goals can be identified, we trained different classification models on the data produced by the human subject study and evaluated the models using cross-validation. The following sections describe the dataset which was used to train our classification models, the feature engineering process and the evaluation of the trained classification models.

3.1 Manually Labeled Dataset

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. This resulted in a

manually labeled dataset containing a total of 2902 instances that was used for training the automatic classification approach.

3.2 Automatic Binary Classification

Our approach for classifying queries consists of two basic steps: Section 3.2.1 describes the POS tagging step, and Section 3.2.2 and 3.2.3 describes details of our classification approach. Throughout our work we used the data mining toolkit WEKA [33] for feature pre-processing, feature selection, classification and evaluation of classification models.

3.2.1 POS Tagging

We can speculate 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[2] for part-of-speech tagging all queries. Due to the lack of POS taggers trained on search query data (which differs from natural language text), our tagger was trained on sections 0 to 18 of the Wall Street Journal part of the Penn Treebank corpus. 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.

3.2.2 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 removed stopwords.

Other feature types we have looked at in preliminary investigations, such as query length, click-through and language modeling, did not exhibit sufficient discriminative power for the task at hand.

In order determine discriminative features, we ranked the features according to chi-square feature selection. Table 1 lists the 20 most discriminative features together with example queries for each feature and the number of occurrences of the feature in the positive class (#). We can observe a change in the ratio between part-of-speech and stemmed word features when considering a larger feature set size. Only a fifth of the features in Table 2 are words, notably all of them verbs. The ratio drifts in favor of the word features when more features are used. Based on the feature selection results, it appears that the most discriminative features for identifying queries containing explicit goals are POS features complemented by verbs.

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html accessed June 3rd, 2008.

Nr.	POS	SOW	#	Feature	Example Matching Query
1	X		126	$ WRB TO	[*$ where to*] find shrooms in Georgia
2	X		130	WRB TO VB	[*how to live*] jewishly
3	X		83	TO VB NN	drink milk [*to lose weight*]
4		X	41	Buy	*buy* now pay later jewelry
5	X		58	VB NN NN	[*find property values*] calculator
6		X	20	Find	*find* an old friend for free
7	X		36	TO VB JJ	I want [*to download instant*] messenger
8		X	27	Make	*make* your own parable
9	X		52	$ VB NN	[$ *find lawyer*] in Georgia to form llc
10	X		29	VB NN IN	[*borrow money from*] Donald Trump
11	X		12	TO VB PRP	how [*to copyright your*] photos
12	X		14	WRB VBP PRP	my hair turned orange [*how do I*] fix it
13	X		26	TO VB NNS	what [*to pay Mexicans*]
14	X		28	VB NN NNS	[*make business cards*]
15	X		19	TO VB DT	teach yourself [*to play the*] piano
16	X		9	VB PRP JJ	how to [*get yourself sick*]
17	X		45	TO VB IN	places [*to stay in*] Gatlinburg
18		X	8	Install	*install* Microsoft windows 2000
19	X		14	$ VB PRP	[*$ customize your*] aol buddy icon
20	X		22	VB PRP NN	how to [*obtain us passport*]

Table 2: Top 20 most discriminative features and example queries that match these features

3.2.3 Classification Methods

We chose a Naive Bayes (NB) classifier as a baseline method [8], and a linear Support Vector Machine (SVM) [6] as a second model.

3.3 Evaluation

In line with prior work on query classification [16], we choose F1, the equally weighted harmonic mean of precision and recall, for evaluation. In conducting experiments with regard to the F1 score, we aimed to identify those configurations that balance precision and recall in a way that is useful for acquiring knowledge about goals.

The two linear classification models were evaluated in the light of varying feature set sizes. In order to generate feature sets of different sizes, we applied chi-square feature selection and kept the top N features for comparison. For each classifier/feature set size combination, 10 trials of three-fold cross-validation were carried out. The resulting F1 score for each combination was averaged over all trials. Figure 2 presents the F1 score of different feature set sizes and classification models.

The results imply that the SVM outperforms the NB classifier in our task, especially when the number of features used increases. To identify explicit goals, the NB classifier depends on accurate feature selection prior to training and classification.

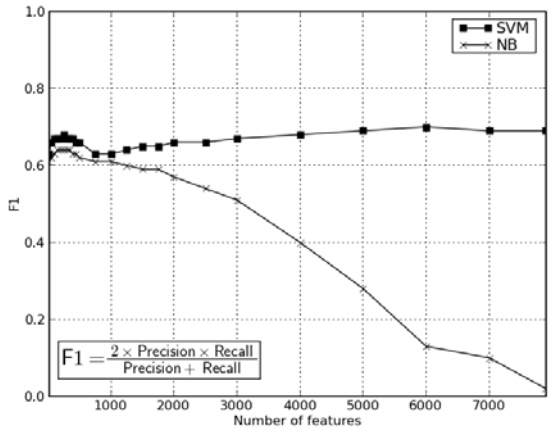


Figure 2: Learn curves for the linear Support Vector Machine and the Naive Bayes classifier. F1 measures with corresponding feature set sizes are displayed

The performance significantly deteriorates when a higher number of features is used. Considering these experimental results, we selected the linear SVM as the classification model for subsequent investigations. In the following, we report the results of the SVM in more detail. All scores are averaged scores resulting from a three-fold cross-validation without feature selection. Table 3 presents the confusion matrix and corresponding True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) scores. The total size of the training set was 2902.

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)

Table 3: The confusion matrix for the linear SVM (using all features) on the manually labeled dataset.

Table 4 shows the precision, recall and F1 scores of our approach. All values refer to the class that represents queries containing goals. A precision of 77% means that in 77% of cases, our classification approach agrees with the majority of human subjects. A simple baseline approach, where we guess that a query containing a verb always contains an explicit goal would perform significantly worse. While the baseline would excel on recall (Recall 1.0, due to our definition of explicit goals that requires a query to contain a verb), it would perform worse on precision (Precision 0.13, based on the data for the manually labeled dataset) and F1 (0.23) scores.

Precision	Recall	F1 – measure
0.77	0.63	0.69

Table 4: Precision, Recall and F1 for our approach (SVM using all features) on the manually labeled data set.

Although there are slight differences in the evaluation procedure and the type of knowledge captured (also see Section 5.2), the precision of explicit goals acquired with our approach is roughly comparable to precision scores reported for the ConceptNet commonsense knowledge database (75%) [28]. These encouraging results suggest the relevance of query logs for this task.

In the following experiment, we use the linear Support Vector Machine trained on the entire feature set without feature selection.

4. EXPERIMENTAL SETUP

Our knowledge acquisition experiment is based on a large search query log recorded by AOL in early 2006. The following section describes the dataset and the various sanitization and pre-processing steps performed.

4.1 Data

We used the AOL search query log [23] since it provides a very large dataset including comprehensive information about anonymous user IDs, time stamps, search queries, and click-through events. It contains ~ 20 million search queries collected from 657,426 unique user ID's between March 1, 2006 and May 31, 2006 by AOL. To our knowledge, the AOL search query log is the most recent, very large corpus of search queries publicly available (2006).

A single query in the AOL log is represented by the following data structure: {UserID, Query, Timestamp, (ItemRank, URL)*}.

4.2 Sanitization & Pre-Processing

Search query logs contain considerable noise which has implications for knowledge acquisition. Types of noise that can be found in the AOL search query log include empty queries, corrupted character encoding and misspellings. In order to reduce noise to an acceptable level, we performed the following sanitization and pre-processing steps: **Empty Queries**: We removed blank queries and queries containing just a minus character. **Short Queries**: Due to the inherent ambiguity of short queries, we restrict ourselves to queries with at least three tokens. This results in a removal of 65% of the queries contained in the original dataset. **URL queries:** We removed queries containing URLs or fragments of URLs using regular expressions. **Queries containing lyrics or movie titles:** In preliminary experiments we observed that queries for music lyrics ("i need love lyrics") often contain a verb, but refer to songs rather than actual user goals. This bears the risk of confusing a classification approach that is in part based on syntactic features. However, such queries can be identified, since they often contain keywords such as "lyrics" or result in click-through to lyrics or movie related websites (e.g. "http://www.seeklyrics.com"). We performed limited term and website blacklisting to heuristically reduce the number of such queries in the dataset. **Syntax check:** We removed queries containing tokens, which are not numbers or sequences of letters. We used this filter to eliminate corrupted character encodings. **Corrected misspellings:** We removed misspelled queries. Whether or not a consecutive query represents a spelling correction was determined by the Levenshtein distance between two concsecutive query strings. A query was removed if the Levenshtein distance between the query and its successor is <= 2 and the first query has no click-through event attached. **Verb filter:** Based on our definition of explicit goals, all queries not containing a verb (as indicated by the POS tagging procedure) were removed. Table 5 summarizes the pre-

processing steps and shows the number of query instances that passed the various filter steps.

Pre-processing step	# of queries
Total number of queries	21.011.038
Empty queries	20.527.902
Short queries	7.242.610
URL queries	6.631.084
Syntax check	5.880.900
Queries containing lyrics or movie titles	5.754.994
Corrected misspellings	5.405.547
Verb filter	1.002.861

Table 5: Pre-processing filter pipeline

Thus, the pre-processed dataset contains 1.002.861 queries. This set also represents the basis for the human subject study reported earlier.

5. RESULTS

In this section, we examine the results produced by applying our automatic classification approach to the pre-processed dataset. Section 5.1 presents selected statistics of the result set and Section 5.2 gives some qualitative insights into the nature of acquired goals.

5.1 Selected Statistics

Applying our automatic classification method to the AOL search query log yielded a result set comprising 118.420 queries, 97.454 of them unique. With a precision of 77%, the result set comprises an estimated *75.039 queries containing explicit goals*, which might appear small in the light of ~20 million queries contained in the original AOL search query log. However, considering that the 20 million queries reportedly represent only 0.33% of the total number of queries served by AOL during that time, the approach would be able to acquire a much larger set of explicit goals on larger datasets.

The 20 most frequent queries from the result set are presented in Table 6. 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".

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

Table 6: The 20 most frequent queries in the result set.

The information in this table could be expected to reflect – to some extent - the needs and motivations of the general public. 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" likely represent false positives, revealing two kinds of shortcomings of our approach: First, the automatic classification approach relies on linguistic patterns generated by part-of-speech tagging. In case of the query "wedding cake toppers", the POS tagger mistakenly tagged "wedding" as a verb (VBG) which results in our classifier producing a false positive. A part-of-speech tagger that is trained on a more suitable corpus might help alleviating such problems in the future. Second, certain queries containing explicit goals resemble the title of books or TV shows, such as "skating with celebrities". This problem could be addressed by including domain knowledge (such as imdb.com) or inspecting and analyzing click-through data and anchor text, which can be expected to improve the overall performance of our approach.

	home (2512)	**card** (2188)	**name** (1844)	**screen** (1561)	**credit** (1433)	**music** (1398)	**money** (1371)	**weight** (1338)	**school** (1221)	**car** (1189)
make (8763)	210	208	96	96	5	58	631	19	19	32
buy (8557)	237	117	12	10	66	58	43	6	17	224
find (8545)	169	25	192	30	20	57	60	17	104	94
get (6562)	65	103	41	26	130	33	68	13	55	54
do (6391)	70	62	72	69	40	51	52	52	44	25
listen (2485)	18	0	0	0	0	477	0	0	27	2
learn (2014)	12	16	3	1	6	34	10	3	28	5
sell (1962)	141	38	8	1	2	8	15	1	1	90
use (1688)	15	22	5	5	15	3	3	10	9	15
play (1598)	8	63	0	1	1	13	3	1	4	4

Table 7: The 10 most frequent verbs and nouns in the result set and corresponding co-occurrences in queries containing goals.

To better understand the nature of the identified goals, we conducted a term analysis by identifying the 10 most frequent nouns and verbs in the result set and analyzing verb/noun co-occurrences. In Table 7 and Table 8, we applied the Porter stemming algorithm for word conflation. Afterwards the stems were manually extended to their base form.

The most popular verb/noun co-occurrences in Table 7 seem to be indicative of typical user goals on the web, such as "make money", listen music" or "buy home". Overall, the top 10 verb/noun correlations identified were "lose weight"* (688), "make money"* (631), "listen music"* (477), "find number"* (457), find address"* (441), "add name" (399), "add screen" (380), "buy online" (339), "find phone"* (333), "find people" (332). Evaluations of the top verb/noun correlations have revealed that many of these goals are also contained in the ConceptNet commonsense knowledge base v2.1 (marked with an *). This can be understood as a further indicator of the usefulness of search query logs for acquiring knowledge about goals. It also suggests that search query logs might be useful to automatically expand the knowledge contained in existing knowledge bases, which has been attempted before [7].

If search query logs would be utilized for such a purpose, 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 6. In Figure 3, goals are plotted according to their rank and the set of different users who share them.

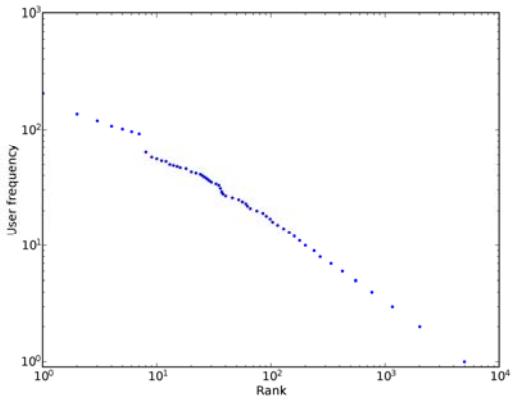


Figure 3: Rank-Frequency Plot of Queries containing unique user goals.

The distribution in Figure 3 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 (similar to the long tail of products [2]). This means that the explicit goals in the result set are extremely diverse; making search query logs a particularly promising resource for the acquisition of *broad* knowledge about user goals in the context of the web.

When we turn our attention to the URLs that have been clicked as a result of performing a query containing a goal, it is interesting to ask, how the rank of websites (the number of times they have been clicked) is affected by our classification approach.

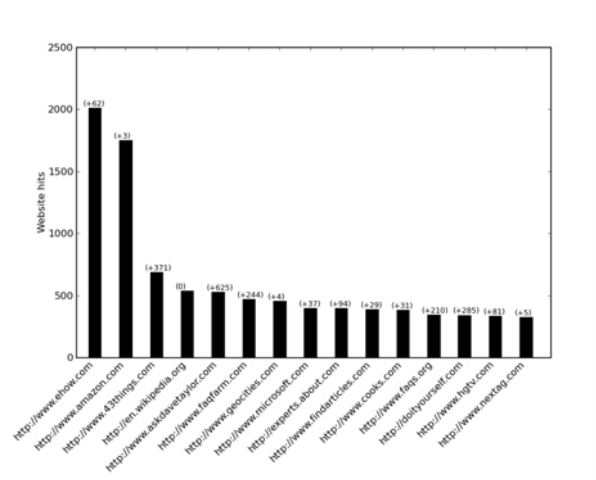


Figure 4: Top 20 websites in the result set according to website hits.

Figure 4 depicts the top 10 websites that have been clicked in our result set. The frequency information is accompanied with the gain/loss in terms of rank compared to the ranking of the websites in the original search query log. The fact that 43things.com – a website aiming to collect user goals – is ranked very high (at position 3) together with a gain in rank of 371 positions suggests that the core ideas of our approach are useful to identify queries containing goals. Other websites affected by our classification approach are websites that can be considered to be very goal-centric: ehow.com (a website on how to accomplish a broad variety of tasks and goals), doityourself.com and hgtv.com (both home improvement websites), faqfarm.com and faqs.com (question answering websites), and medhelp.org (a medical information website). Medhelp.org is a particularly interesting result of a website affected by our approach (ranked 33, not depicted in Figure 4), as a large proportion of the queries containing goals are queries describing medical symptoms ("coughing up clear fluid"), which we defined as avoid goals.

5.2 Qualitative Analysis

While the analysis conducted so far gave statistical insights into the nature of goals contained in search queries, it is difficult to infer information about their quality. To address this issue, we have performed limited qualitative analysis through inspection. We selected an arbitrary set of verbs and nouns and inspected the acquired goals related to these verbs or nouns.

In Table 8, the 10 most frequent goals in the result set are listed, which contain either of the verbs "get", "make", "change" or "be". Frequency refers to the occurrence of the goal in the result set. The goals listed in Table 8 are the result of identifying the first verb in a query containing a goal, and truncating any tokens prior to this verb. Again, queries marked with a "*" represent queries that are contained in ConceptNet's commonsense knowledge base (v2.1) as well. Many verb phrases in Table 8 are related to existing commonsense knowledge goals, such as "be pregnant", "be rich" or "be funny".

Nr.	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)

Table 8: The 10 most frequent verb phrases containing the verbs "get", "make", "change", and "be" are listed.

To gain further insights, we focus on sets of related goals from four arbitrary categories: "Money" [economy], "Birth" [event], "Home" [location] and "People" [person]. Some heuristic post-processing was applied to the queries: "How to" was replaced by "Learn to" and "where to" by "find location to". The generated goal frequency plots are shown in Figure 5.

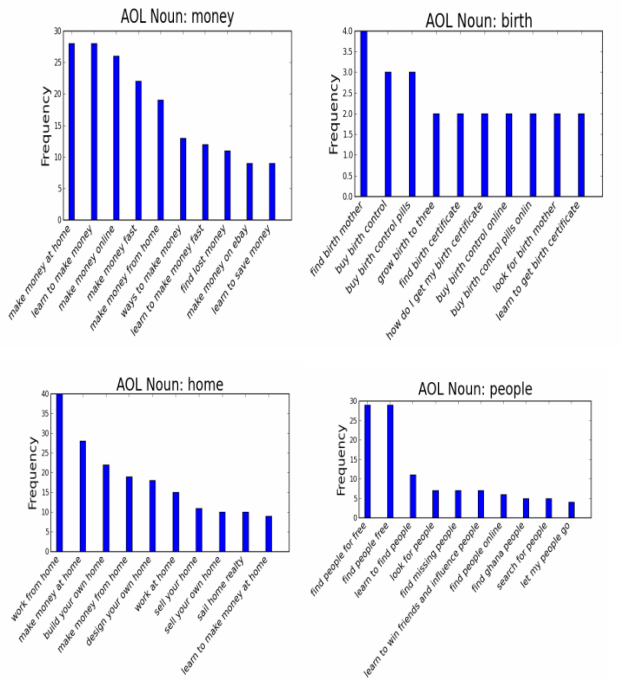


Figure 5: Explicit goals that contain the nouns "money", "birth", "home" and "people".

These plots make an interesting case for using search query logs for expanding existing knowledge bases. Queries containing goals could be selected based on their semantic similarity to existing nodes in a knowledge base ("e.g. "money"). The set of related goals produced for a given node hint towards the potential usefulness of search query logs for acquiring/expanding knowledge about user goals on the web.

6. COMPARISON TO RELATED WORK

In previous research, He et al. [10] 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 novel in the sense that it studies search queries themselves as a source of explicit goals, which can be suspected to better reflect user intent. In the context of user intentions and query log analysis, search queries have been the subject of research for several years: An influential study conducted by Broder in 2000 [4] introduced a high level taxonomy of search intent, proposing a distinction between three classes of search goals: navigational, informational and transactional queries. This has stimulated a series of follow-up research on category refinement and automatic query categorization [3][12][14][25]. While previous research in this area has achieved considerable progress in the *categorization of queries* into high-level goal taxonomies serving a primarily *functional purpose* (to improve search, cf. [3][12][14]), we know little about the *acquisition of goal instances (explicit goals)* from search query logs for *knowledge capture* purposes (as in [18][20]).

Beyond related research in query log analysis [4][12][14][22], our work is relevant to other research areas, notably commonsense knowledge acquisition and reasoning, goal mining and intelligent user interfaces.

Examples of related research in the area of *Commonsense Knowledge* are Openmind Commonsense and the related ConceptNet project, which tap into "knowledge contributions" made by volunteers on the web. ConceptNet contains a wide variety of commonsense knowledge including knowledge about human motivations and intentions [20]. ConceptNet's relation types "MotivationOf" and "DesireOf" are examples of attempts aiming to capture different aspects of intentional commonsense knowledge. In addition to volunteer-based knowledge acquisition, search engine result snippets have been studied for that purpose as well [7], and the idea of "human computation" [32] has inspired the development of games for collecting commonsense goals [18].

Goal Mining is often referred to as the acquisition of goals from textual resources. This research area covers a broad range of interesting problems, including the acquisition of goals from patents, scientific articles [11], organizational policies [24], organizational guidelines and procedures [17] and others. Acquiring explicit goals *from search queries* represents a unique problem in the context of Goal Mining, and – to the best of our knowledge – has not been studied before.

In the context of *User Interfaces*, goals can give answers to the "why" questions about user behavior and user interactions [18] and have been found to play a fundamental role in explanation, justification, and rationalization. In commonsense enabled applications, explicit representations of goal knowledge are crucial for plan recognition and planning [18], and are an enabler for intelligent user interfaces that exhibit traits of common sense understanding, such as goal-oriented search [19] or goal-oriented event planning [29].

7. DISCUSSION

7.1 Threats to Validity

In the following, we describe threats to validity of our results [34]:

Construct validity: The main construct we investigate in this research is the notion of *explicit goals contained in search queries*. While our definition intentionally gives some room for variability, our human subject study yielded reasonable scores for inter-rater agreement κ and a reasonable distribution of e(q), which can be interpreted as preliminary empirical evidence for the agreeability of our construct.

Internal validity: The subjects involved in our human subject study were graduate students enrolled at Austrian universities, who were not involved in the research of this paper. While all subjects were fluent in English, they did not share the same cultural context of the population that performed the queries contained in the AOL search query log (Austrian Graduate students vs. North-American AOL users). Although we could not find evidence of this problem in our human subject study, it could be the case that certain queries were mistakenly labeled as containing an explicit goal (such as "saving private ryan"). However, such queries could be easily eliminated using domain knowledge in future work (e.g. imdb.com). Our bias towards longer queries (n>2) prohibits us to study a large part of the AOL search query log (~65%). The focus on longer queries was motivated by the inherent difficulty of part-of-speech tagging one or two word English queries correctly, and by the fact that search engine vendors report increasing average query lengths over the past years[3].

External validity: In our definition of explicit goals, we are referring to existing work on intentions and goals from different research areas including human-computer interaction, goal-oriented requirements engineering [26] and search query analysis [31]. We have found that the explicit goals contained in our result set are biased towards the population contributing to the AOL search query log. This is reflected in popular goals such as "cancel AOL account" or in popular events during the time of data collection, such as "betting on american idol". However, we believe that our approach is general enough to be successfully applied to other search query logs as well.

Reliability: In our investigations, we have used existing toolkits such as the WEKA toolkit and established algorithms, so that reproducing our results is possible. More details regarding the implementation of our approach will be made available on a website accompanying this paper.

8. CONCLUSIONS

Our work shows that search query logs have the potential to address two fundamental 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. The automatic classification approach presented demonstrates that goals can be acquired from search query logs with useful precision/recall scores, showing significant improvements over a simple baseline approach. Because query logs are a natural byproduct of human activity on the web, the costs associated with knowledge acquisition can be expected to be lower compared to other approaches requiring, for example, knowledge engineers or volunteers. Quantitative and qualitative inspections of our results revealed that the goals acquired from search query logs in part represent commonsense knowledge and cover a vast range of topics and levels of granularity, which makes search query logs an interesting resource for addressing the goal coverage problem. Classifying these goals into existing taxonomies of goals and search intent [4] in future work might yield new insights into the motivations and goals of users on the web. Ongoing work focuses on the problem of correlating queries containing explicit goals ("buy a car") with each other, and with queries that are not containing explicit goals ("car dealership") to address the construction of conceptual nets representing and correlating knowledge about human goals.

[3] http://blogs.zdnet.com/micro-markets/index.php?p=27, last accessed June 3, 2008

9. REFERENCES

[1] J. Allan and H. Raghavan. Using part-of-speech patterns to reduce query ambiguity. Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 307--314, ACM Press New York, NY, USA,2002.

[2] C. Anderson. The Long Tail. Wired Magazine, (12)10, 2004.

[3] R.A. Baeza-Yates, L. Calderon-Benavides and C.N. Gonzalez-Caro. The intention behind web queries. In F. Crestani and P. Ferragina and M. Sanderson, editor(s), Proceedings of String Processing and Information Retrieval (SPIRE), (4209):98--109, Springer, 2006.

[4] A. Broder, A taxonomy of web search, SIGIR Forum, vol. 36, no. 2, pp. 3-10, 2002.

[5] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, (20)1:37, 1960.

[6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In Proceedings of the International Conference on Information and Knowledge Management (CIKM'98), New York, NY, USA, ACM Press, pp 148-155, 1998.

[7] I. S. Eslick. Searching for commonsense. Masters Thesis, MIT, 2006.

[8] J.H.Friedman. On bias, variance, 0/1loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery. Kluwer Academic Publishers, 55-77, 1997.

[9] L. Gravano, V. Hatzivassiloglou and R. Lichtenstein. Categorizing web queries according to geographical locality. In Proceedings of the International Conference on Information and Knowledge Management (CIKM'03), 325--333, ACM,New York, NY, USA, 2003.

[10] 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 2007

IEEE/WIC/ACM International Conference on Web Intelligence, 683-686, IEEE Computer Society, 2007.

[11] B. Hui and E. Yu. Extracting conceptual relationships from specialized documents. Data & Knowledge Engineering, (54)1:29-55, Elsevier, 2005.

[12] 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.

[13] D. Kirsh. When is information explicitly represented? Information, Language and Cognition - The Vancouver Studies in Cognitive Science., 340-365, UBC Press,1990.

[14] U. Lee, Z. Liu and J. Cho. Automatic identification of user goals in Web search. In Proceedings of the 14th International World Wide Web Conference (WWW'05), 391-400, ACM Press,New York, NY, USA,2005.

[15] D.B. Lenat. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, (38)11:33-38, 1995.

[16] Y. Li, Z. Zheng and H.K. Dai. KDD CUP-2005 report: facing a great challenge. ACM SIGKDD Explorations Newsletter, (7)2:91--99, ACM Press New York, NY, USA, 2005.

[17] S. Liaskos, A. Lapouchnian, Y. Yu, E. Yu and J. Mylopoulos. On goal-based variability acquisition and analysis. In Proceedings of the 14th IEEE International Requirements Engineering Conference (RE'06), Minneapolis, USA, 2006.

[18] H. Lieberman, D.A. Smith and A. Teeters. Common Consensus: a web-based game for collecting commonsense goals. In Proceedings of the Workshop on Common Sense and Intelligent User Interfaces held in conjunction with the 2007 International Conference on Intelligent User Interfaces (IUI 2007), 2007.

[19] H. Liu, H. Lieberman and T. Selker. GOOSE: A goal-oriented search engine with commonsense. AH '02: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 253--263, Springer-Verlag,London, UK,2002.

[20] H. Liu and P. Singh. ConceptNet - A practical commonsense reasoning tool-kit. BT Technology Journal, (22)4:211-226, 2004.

[21] G.C. Murray and J. Teevan. WWW Workshop Report: Query log analysis - social and technological challenges. ACM SIG IR Forum, (41)2 ACM New York, NY, USA, 2007.

[22] M. Pasca, B. Van Durme and N. Garera. The role of documents vs. queries in extracting class attributes from text. In Proceedings of the International Conference on Information and Knowledge Management (CIKM'07), 485-494, ACM, New York, NY, USA, 2007.

[23] G. Pass, A. Chowdhury and C. Torgeson. A picture of search. Proceedings of the 1st International Conference on Scalable Information Systems, ACM Press New York, NY, USA, 2006.

[24] C. Potts, K. Takahashi and A. I. Anton. Inquiry-based requirements analysis. IEEE Software, (11)2, 1994.

[25] D. E. Rose and D. Levinson, Understanding user goals in web search. In Proc. of WWW 2004, May 17-22, 2004, New York, USA, 2004.

[26] G. Regev and A. Wegmann. Where do goals Come from: the underlying principles of goal-oriented requirements engineering. RE '05: Proceedings of the 13th IEEE International Conference on Requirements Engineering (RE'05), 253--362, IEEE Computer Society, Washington, DC, USA, 2005.

[27] R.C. Schank and R.P. Abelson. Scripts, plans, goals, and understanding: an inquiry into human knowledge structures. Lawrence Erlbaum Associates, 1977.

[28] P. Singh, T. Lin, E.T. Mueller, G. Lim, T. Perkins and W.L. Zhu. Open Mind Common Sense: Knowledge acquisition from the general public. In Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, 1223--1237, Springer-Verlag London, UK, 2002.

[29] D.A. Smith. EventMinder: A Personal Calendar Assistant That Understands Events. Master Thesis, MIT, 2007.

[30] M. Strohmaier, M. Lux, M. Granitzer, P. Scheir, S. Liaskos and E. Yu. How do users express goals on the web? - an exploration of intentional structures in web search. We Know'07 International Workshop on Collaborative Knowledge Management for Web Information Systems, in conjunction with WISE'07, Nancy, France, 2007.

[31] M. Strohmaier, P. Prettenhofer and M. Lux. Different degrees of explicitness in intentional artifacts - studying user goals in a large search query log. Proceedings of the CSKGOI'08 Workshop on Commonsense Knowledge and Goal Oriented Interfaces, held in conjunction with IUI'08, Canary Islands, Spain, 2008.

[32] L. von Ahn. Games with a purpose. Computer, (39)6:92--94, IEEE Computer Society Press Los Alamitos, CA, USA, 2006.

[33] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd ed., ser. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, June 2005.

[34] R. K. Yin, Case study research: design and methods (Applied Social Research Methods). SAGE Publications, December 2002.