Ensembling Classifiers for Detecting User's Intentions behind Web Queries

Alejandro Figueroa[1,2] and John Atkinson[3]

(1) Yahoo! Research, Santiago, Chile, afiguero@yahoo-inc.com

(2) School of Informatic Engineering, Universidad Diego Portales, Santiago, Chile,

alejandro.figueroa@mail.udp.cl

(3) Department of Computer Sciences, Faculty of Engineering, Universidad de Concepcion,

Concepcion, Chile, atkinson@inf.udec.cl

Abstract

Discovering a user's intentions behind search queries is a key issue to improve users experience by adapting results to their goal. Thus, automatically detecting a user's intention when searching is at the core of successful information retrieval systems on the Web. The task can usually be seen as a classification problem in which a sample of annotated user query intentions are provided to a supervised machine learning algorithm or classifier, which learns from these examples, and then it can classify unseen user queries. In order to deal with these issues, in this paper a new approach based on an ensemble of classifiers is proposed. It combines syntactic and semantic features so as to effectively detect user's intentions. Different setting experiments show the promise of our linguistically-motivated ensembling approach by reducing the ranking variance of single classifiers across user's intentions.

Index Terms

Machine Learning, Ensemble Learning, Natural-Language Processing, Web Search.

I. Introduction

In the last years, the web has not only become a huge repository of information, but also a place where people can interact and access different kinds of resources such as services and applications. However, there is a gap between user needs and the resources to meet them. Users express their requests by entering a short sequence of query terms, which are further interpreted by search engines in order to provide relevant answers. This makes search engines key players in understanding and efficiently resolving hundreds of millions of queries per day.

In order to get valid query interpretations, a main step involves discriminating the user's intention, which varies from fulfilling information needs to using search engines as navigational tools to reach specific web sites. Search

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

engines can also be used to perform transactions by providing access to different types of resources including maps, lyrics, and books. Automatically detecting user's intentions is a key challenge for search engines as they can improve user's experience by getting more useful results and tailoring them to their specific needs. On the one hand, the intention of some highly frequent queries (e.g. "*wikipedia*" and "*yahoo*") can easily be identified by benefiting from any type of hash table extracted from analyzing click patterns across search logs. Furthermore, a user's intention behind queries with a limited set of patterns (i.e., *term1 term2 lyrics* and *define term1 term2*), can also be readily recognized. Nevertheless, it is hard to determine the intention of a large portion of new queries by using simple heuristic patterns.

Thus, automatically detecting a user's intention when searching is at the core of successful information retrieval systems on the Web. This task can usually be seen as a supervised learning problem (i.e., classification) in which word-based learning algorithms (i.e., classifiers) search through a hypothesis space to find a suitable hypothesis that will make good predictions with a intentions detection problem [11], [14]. Even if the hypothesis space contains hypotheses that are very well-suited for a detection task, it may be very difficult to find a good one.

In order to address similar tasks, methods ensembling multiple classifiers have started to catch the attention of the research community in the last ten years [13]. Several strategies have been designed for tackling distinct problems, for instance, for semantically classification of search queries [16], [15].

In this paper, a novel approach based on an ensemble of classifiers is proposed. Unlike previous approaches, our research takes advantage of a specific type of ensembles via classifier selection so as to improve the recognition of the user's intent behind search queries. The model combines syntactic and semantic features so as to effectively detect a user's intention using different ensembling techniques for detecting a user's intentions [3], [17].

II. Related Work

Some studies have proposed a taxonomy for web search engine queries based on manual inspections [4]. A first level consists of three canonical branches, which cover most of user's goals when searching: **navigational** (e.g., "*facebook*" and "*twitter*"), **information-oriented** (e.g., "*how do I get rid of acne?*" and "*obama bio*"), and **resource-oriented** (e.g., "*berlin map*" and "*free anti-virus*").

Current approaches to automatically labeling/classifying search queries [9] randomly select instances extracted from search logs. They aim at discovering relevant features to discriminate one intention from the other, which included keywords and information extracted from the pages visited by users. Resulting navigational queries were found to be classified by organization and people names (e.g., "*dell*" and "*madonna*"), and domain suffixes (e.g., "*.com*"). On the other hand, resource queries are short and likely to contain keywords such as lyrics, movies, recipes, and images (e.g., "*lentil soup recipe*" and "*justin bieber images*"), whereas informational queries are longer, and

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

usually formulated with question words resembling natural language text (e.g., "*What is the biggest organ in the human body?*").

Other methods group web queries based on these three canonical segments using k-means clustering and a feature-rich representation. Each item in the search log comprises features such as user identification, cookie, time of day, query terms, and the type of content collection the user is searching for. In addition, each item was enriched with the query length, a number modeling the search engine results page visited during a given interaction, the number of times a user changed the query during a session. The method then assigns terms to each record such as as informational, navigational, or transactional [9].

Statistical language models have also been exploited to classify web query instances based on their intention [9]. These instances are then used to automatically categorize new queries via exact terms matching. However, the approach is too restrictive as it matches frequent elements. In order to deal with this issue, intention classification approaches use Support Vector Machines (SVM) and Naive Bayes classifiers [8], showing that a SVM obtained better results on the informational category whereas Naive Bayes did well for the other two types of intentions. Experiments indicate that word-based features become key to recognize resource queries, but they perform poorly on the navigational class.

A recent work studied the linguistic difference between search queries and text documents [2], discovering that ca. 70% of query terms are nouns and proper nouns, whereas adjectives are used ca. 7% of the time and URLs 6%. As for documents, almost each sentence contained at least one verb. Since this poses a great challenge to conventional natural-language processing techniques, new ad-hoc algorithms have been designed for dealing with search queries in order to assist in detecting the user intention by using *Named-Entity Recognition* (NER) techniques [1], [5], [7], [18], [6].

III. Ensembling Classifiers for Detecting User Intentions

Classification or supervised learning is a machine learning task of inferring a function from labeled training data. The training data consist of a set of labeled examples which are pairs consisting of an input object and a desired output value. Whereas *ensemble learning* refers to a collection of classification methods that learn a target function by training a number of single classifers and combining their predictions. The principle is that a committee decision, with individual predictions combined appropriately, should have better overall accuracy, on average, than any individual committee member. For many tasks, ensemble models often attain higher accuracy than single models as a more reliable function-sample mapping can be obtained by combining the output of multiple 'experts'.

Accordingly, this work addresses automatic recognition of user's intentions behind web queries by extending and ensembling current classification models so as to improve search experience. Instead of focusing on single classifiers for detecting different types of intentions [8], in this research, ensembles of single classifiers are explored.

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

In order to deal with the user intention detection using multiple types of queries, our fusion considers supervised single classifiers that can relatively easy cope with multi-class problems:

1) *Multi-Class Support Vector Machines (SVM)*[1] are kernel-based supervised learning models with associated learning algorithms that classify data into several categories [12].
2) *Naïve Bayes Classifier*[2] is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
3) *Maximum Entropy (MaxEnt) classifier*[3] is a probabilistic classifier based on the Principle of Maximum Entropy which assumes the features to be conditionally independent. MaxEnt selects the model fitting the training data which has the largest entropy.
4) *Multi-Layer Perceptron (MLP)*[4] is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. MLP uses a supervised learning technique called *Backpropagation* for training the network.

In order to train these learners (classifiers), a corpus of web queries from the AOL[5] query collection was captured and annotated. Three types of ensembles' classifiers were then constructed using several specific-purpose features:

1) A **General Ensemble** combines the four single classifers to produce a final (intent) class label. Each classifer is general so that it has access to all query set, when building its classification model.
2) A **Syntax-oriented Ensemble** uses a syntactic-based taxonomy (i.e., a classification of things or concepts, including the principles that underlie such classification) of search queries [2], by splitting each general single classifer into four single classifers which focus on a specific type of query.
3) A **Length-oriented Ensemble** creates different classifiers targeted at web queries of distinct lengths. Unlike the previous ensambles, each single classifer is split into single classifers based on the number of terms in the search query, so five groups were generated. Accordingly, each single classifer has access only to the respective group of queries, when constructing its classification model.

A. Corpus Acquisition

In order to prepare a dataset for detecting a user's intention, the AOL web query corpus[6] was used which contains 21 million query instances submitted by approximately 650 thousands of search users. Each instance contains a

[1]http://svmlight.joachims.org/svm_multiclass.html
[2]http://mallet.cs.umass.edu
[3]http://www.cs.cmu.edu/∼aberger/maxent.html
[4]http://www.cimne.com/flood/
[5]http://www.infochimps.com/datasets/aol-search-data
[6]http://gregsadetsky.com/aol-data/

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

user id, timestamp, the query string, the rank and the URL of the results clicked by the user. The overall collection consists of about ten millions of distinct lowercased queries, where 4,811,638 elements are linked with at least one clicked URL. These queries were then **mannually annotated**, first extracting a sample of 30,000 random queries from the remaining 3,788,459 unlabeled items, which were annotated by ten human annotators. Each annotator was provided with an set of 3,000 distinct queries and the description of each class of intention so as to reduce ambiguity when tagging. As a result, manually annotated categories included 23,736 informational, 4,585 navigational, and 1,679 resource queries.

Since each annotator was assigned a different set of queries, the disagreement rate was approximated by inspecting the annotations for one hundred random instances tagged by each of them. The lowest and the highest disagreement were 2% and 15%, respectively, with an average of 8% (std. deviation of 4.37).

Despite the preliminary annotation results, manually tagging is a time demanding task, and may be biased by human criteria. In addition, intentions have been observed not to be that ambiguous for humans so there is no need for too many annotators. Hence manual annotation was combined with rule-based **automatic annotation**. For this, some rules were defined to include common words that unambiguously imply a particular class of query intention [9].

By using these rules, only 21.26% of the queries were tagged, whereas other elements required manual annotations. From these tagged samples, 300 randomly selected instances were manually inspected (error rate of 7.33%). Overall, 1,023,179 items were automatically extracted (793,314 navigational, 122,805 resource and 107,060 informational)[7].

B. Ensembling Single Learners

In our approach, two ensembling strategies are used:

- *Borda Count Method:* this is a single-winner election method in which voters rank options or candidates in order of preference. The Borda count determines the outcome of a debate by giving each candidate, for each ballot, a number of points corresponding to the number of candidates ranked lower. Once all votes have been counted the candidate with the most points is the winner.
- *MaxEnt Ensemble:* Maximum Entropy ensembling models possess several desirable features such as flexibility in adding new features, scalable training, easy parameter estimation and minimal assumptions about the posteriors. In our approach, Maximum Entropy models are trained with two kinds of features: the user's intention returned at each ranking position by each learner, and binary features indicating whether two or three classifiers rank the same intention in the same position. A Greedy algorithm was then applied to select the best features for this type of ensemble.

[7]The annotated datasets can be obtained from $http://www.inf.udec.cl/\sim atkinson/AnnotatedCorpus.rar$

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

C. Features for each Single Learner

Unlike other approaches [8], several features were captured so as to to take advantage of each single learner. Overall, five groups of features were identified:

1) Term-level features such as words and the query length.
2) *Caseless Models:* a set of fine-grained features was extracted from caseless corpora by conducting some *Natural-Language Processing (NLP)* [10] tasks including:
   a) *Named-Entity Recognition (NER):* is the task of identifying and classify atomic elements (named entities) in text into pre-defined categories such as the names of persons, organizations, locations, etc.
   b) *Dependency Parsing:* in NLP, parsing or syntactic analysis is the task of analysing a sentence of words into its structure, resulting in a *parse tree* showing their syntactic relation to each other, which may also contain semantic and other information. A typical structure is based on constituents (i.e., noun phrases linked to verb phrases, etc) or dependency relations. A dependency relation views the (finite) verb as the structural center of all clause structure. All other syntactic units (e.g. words) are either directly or indirectly dependent on the verb. Thus, structure is determined by the relation between a word (a head) and its dependents.

   By using state-of-the-art NLP methods and publically available tools[8], extracted features included:

   - **Named Entities** representing organizations, persons, locations.
   - **Dependency relations** were extracted from the queries' **Dependency Trees** by using the Stanford Dependency Parser. Obtained dependency information (dependency paths) included:
     - The number of dependency relations (out of 109 distinct values from 6,000 queries).
     - The total number of dependency relations.
     - Lexical relationships such as full (typed) relations (i.e., $prep\_in : falling \rightarrow love$) or *partial (typed) relations* (i.e., $prep\_in : falling$)
     - Root model features such as the value and the position of the root node.
3) *Named Entities in Queries:* boolean features were added to indicate the presence/absence of 20 different categories of entities distinguished by a NER such as brand names, business, disease and condition, dish, food, place name and product, etc.
4) *Barr's Taxonomy:* query's boolean features representing syntactical information were obtained from a specific-purpose taxonomy [2]. These identify queries that are Noun Phrases (NP), questions, URL, and Verb Phrases (VP).

[8]http://nlp.stanford.edu/software

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

5) *Query Expansion:* it reformulates a query to improve document retrieval performance, and involves evaluating a user's input and expanding the search query to match additional documents. In order to carry out this expansion, several sources of semantic information were exploited:

- **WordNet:** It is a lexical database which groups words into sets of semantic synonym relationships called *synsets*, recording a number of relations among these synonym sets or their members. By using *WordNet*[9], 26 distinct semantic relations were found across our collection including hypernyms (e.g., cover $\rightarrow$ conceal) and meronyms (e.g., motorcycle $\rightarrow$ kick starter).
- **Wikipedia-based features:** six boolean features were included to indicate whether a query was expanded with words contained in five different sections of Wikipedia: *abstracts, first paragraphs, categories, types of infoboxes*, and *sense discriminators*. In order to look-up articles, a case insensitive match between the title and the search query is looked for, and a $Freebase$ category related to that matched article was included. In order to map wikipedia articles into Freebase categories, the WEX collaboratively-edited legal dictionary[10] was used.

Features for each single classifer are selected by using a *greedy* search algorithm which follows the problem solving heuristic of making the locally optimal choice at each stage so as to hopefully find a global optimum (or a global approximate optimal solution in a reasonable time). The algorithm starts with an empty bag of features and after each iteration adds the one that performs the best. In order to determine this feature, this search method tests each non-selected features together with all the features in the bag. The algorithm stops when there is no non-selected features that improve the performance.

IV. Experiments and Results

In order to assess the performance of our multi-class ensembling method, several experiments using different ensembling configurations were conducted, using the *Mean Reciprocal Rank* (MRR) which is a statistic metric for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. Note that this is preferrable to $P@n$ (i.e., Precision at n) as MRR provides much information on the classification errors. The MRR is the average of the reciprocal ranks of results for a sample of queries Q: $\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$. Furthermore, for benchmark purposes, two baselines were considered:

- A *Centroid Vector* (CV) was built for each user's intention. Each testing sample was then assigned the class label corresponding to the best scoring CV. This baseline has an MRR of 0.8376.

[9]http://wordnet.princeton.edu

[10]http://www.law.cornell.edu/wex/faq

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

- Single Classifiers were trained with words as features. The highest MRR score was 0.8675 using MLP, which was set as baseline.

Classifiers were trained and tested for both baselines, and ensembled by using the *Borda Count* and *MaxEnt* methods.

In order to assess our different ensembles (i.e., general-purpose classifiers, syntax-oriented classifiers, and length-oriented classifiers), a $k - fold$ ($k = 10$) cross-validation was performed. Here the original sample is randomly partitioned into k equally sized subsamples, where one is retained as the test data, and the remaining k - 1 subsamples are used for training. This process is then repeated k times (the folds), with each of the k subsamples used exactly once as the test data.

Table III shows an overview of the three configurations for our experiments. In the first place, all ensembles were observed to outperform the use of single models, independently on exploiting an unsupervised approach (Borda Counts) or a supervised approach (MaxEnt). However, the supervised approach finished with better results than the unsupervised counterpart. Accordingly, Borda Counts may be a cost-efficient solution to improve performance. On the other hand, the fact that Configuration II and Configuration III outperformed Configuration I indicates that the classifier selection is a promising contribution to this task. In words, splitting the search query space according to some intrinsic features (e.g., their length or some syntactic pattern) and building classifiers specialized in each of those features improves the overall performance. In our case, computing the number of tokens/terms within the a search query is efficient as it is computationally cheap.

The improvement of the method using classifier selection may be due to that the feature optimization algorithm is capable of filtering those elements that are more suitable to each region. Furthermore, even though some features might be included into several specialized classifiers, the distribution of its values might radically differ from one region to the other. As a natural consequence, focused classifiers can capture these differences across intentions more effectively. Thus, improving the ranking of intention is a key factor to tailor results that fit the display into small modern devices such as tablets and mobile phones.

A. Configuration I: General-Purpose Ensemble

Each general-purpose single classifier outperformed both baselines. The worst single classifier (i.e., MLP) obtained an MRR of 0.8988, which is 3.61% better than the other classifers using words as features (MRR=0.8675). Overall, the best classifier (MaxEnt) outperformed the worst one by 0.76% (MRR=0.9101) suggesting that the set of previously defined features is useful for automatically detecting the user's intention. This may due to that beyond bag-of-words, no other features were incorporated into the four classifiers. However, four features are used into three of the classifiers: *no. of relations* (extracted from the dependency tree using a parser), file names

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

identified by the NER task, query expansion terms contained in the first paragraph of *Wikipedia* articles, and first-level categories provided by WEX. The most discriminative cues were indeed extracted from Wikipedia and WEX: The 'people' category was linked 68% with navigational queries; whereas 32% was linked with informational intentions. Furthermore, the 'music' class was linked 80% with navigational intentions; whereas 20% was linked with informational intentions. In general, getting a first-level WEX category for the query is associated 76% and 24% with navigational and informational queries, respectively. On the other hand, file names were related 83% to resource queries and 17% to informational intentions.

Since both types of ensemble may improve the performance compared with the best single classifier (MaxEnt) with MRR=0.9101: Borda Counts by 0.30% and MaxEnt Ensemble by 0.46%, further experiments assessing other types of ensemble of classifers become promising. Results also show that a supervised approach such as MaxEnt Ensemble, can bring further benefits as key features are associated with top-ranked intentions of each classifier. Experiments also suggest that a SVM outperforms the other classifers when dealing with informational queries (MRR=0.9852), whereas a MLP performs better on the other two intentions ($MRR = 0.9037$ for navigational and $MRR = 0.8583$ for resources queries). Unlike previous work [8], the multi-layer nature of Perceptron classification models are promising when comparing with Bayes classifiers, as they can capture more complex relationships between queries and target classes. Nevertheless, the performance of SVM significantly drops when dealing with resource ($MRR = 0.6226$) and navigational ($MRR = 0.8395$) queries, so the performance of the MLP drops on the informational type as it achieves an MRR of 0.9019 (8.5% lower than for SVM), showing a significant variance across the three intentions for these two learners.

Experiments also indicated that Bayes and MaxEnt get higher MRR scores than the other classifiers, improving the rank. Indeed, MaxEnt Ensemble did not outperformed SVM on the informational class, nor the MLP model on the other two intentions. However, it got the best overall MRR score, which may be due to that weighting the outcomes of multiple classifiers reduces large variances across distinct intentions as compared to single learners.

B. Configuration II: Syntax-oriented Ensemble

For this experiment, the general classifier was split into four single classifers based on a syntactic taxonomy [2], where a single classifer is a classifier focused on a specific kind of query (i.e., a smaller group of queries showing a specific pattern).

As seen in table I, using smaller units assisted MaxEnt and Bayes Classifier to improve classification accuracy with a slightly higher MRR score than previous settings (0.9101→0.9105 and 0.9056→0.9063, respectively). On the other side, for MLP and SVM classifiers, the MRR performance decreased from 0.8988→0.8945 and

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

Single Learner	No. Samples	MLP	MaxEnt	Bayes	SVM	**Ensemble**	
						Borda Counts	MaxEnt
question	2943	0.9929	0.9898	0.9911	0.9893	0.9919	**0.9949**
url	10119	0.9965	0.9966	0.9968	0.9967	0.9967	**0.9973**
noun phrase	24590	0.8553	0.8760	0.8717	0.8270	0.8766	**0.8806**
others	22348	0.8784	0.8991	0.8921	0.8818	0.9031	**0.9066**
Syntax-oriented Learner	60000	0.8945	0.9105	0.9063	0.8840	0.9124	**0.9156**

TABLE I
ENSEMBLING SYNTAX-ORIENTED SINGLE LEARNERS (CONFIGURATION II).

0.8853→0.8840, respectively, suggesting that using all-data encompassing classifiers is a much more cost-efficient than grouping the same classifier into syntactically targeted units.

In terms of syntax-oriented classifiers, MLP, Naive Bayes and MaxEnt classifiers got the best performance for the question type, for the URL group and for the NP groups, respectively. However, all syntax-oriented single classifers achieved a high performance for URL and question groups as these groups are biased towards a particular intent. Furthermore, most of the instances contained in the URL group are navigational (99.37%); whereas they are informational queries in the question group 97.86%. The distribution of intentions across the NP group is similar to that found across the 60,000 queries: 43.31% for informational, 48.68% for navigational, and 8% for resources. However, for the "other" class, 59.71% are informational, whereas 25% are navigational, showing that the intentions of NP are the most difficult to detect, followed by the "other" group. Nevertheless, an ensemble of single classifers improved the performance of the best single classifier (MaxEnt) on the NP group from 0.8760 to 0.8806 (0.53%).

For query intentions, the MLP classifer outperformed the other three classifers when coping with the navigational (0.9032) and resource (0.85) queries, whereas the Bayes classifier was the best on the informational class with an MRR score of 0.9477, mainly due to few features were incorporated into more than one single learners. In particular, for MLP, MaxEnt, Bayes Classifier and SVM only 2, 8, 5 and 2 features respectively were considered into more than a single classifier, respectively, where duplicated features were often due to named entities extracted via NER tasks.

C. Configuration III: Length-oriented Ensemble

For length-oriented single classifiers, the performance was better than for configurations I and II, as seen in table II. The performance of the SVM increased by 2.39% (Configuration I: two-tailed $t-test = 3.92, n = 20, \alpha = 0.01, p < 0.001$), whereas the Naive Bayes classifier ($p < 0.001$), showed a slight growth by 0.30%. The Borda Counts ensemble also improved the MRR score from 0.9128 (Configuration I) to 0.9163. Furthermore, both ensembles outperformed all four length-oriented single learners, where MaxEnt ensemble improving by 0.74% ($t-test = 3.675, n = 20, \alpha = 0.01, p < 0.001$), suggesting that a length-oriented ensemble significantly performs

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

Length	No. Samples	MLP	MaxEnt	Bayes	SVM	Ensemble	
						Borda Counts	MaxEnt
1	15260	0.9680	0.9681	0.9683	0.9681	0.9683	**0.9687**
2	12868	0.8301	0.8414	0.8444	0.8311	0.8467	**0.8554**
3	11874	0.8877	0,9098	0.9066	0.9005	0.9102	**0.9152**
4	8606	0.8987	0.9104	0.9091	0.9073	0.9138	**0.9203**
5+	11392	0.9247	0.9243	0.9277	0.9148	0.9338	**0.9380**
Length-oriented Learner	60000	0.9044	0.9128	0.9133	0.9065	0.9163	**0.9211**

TABLE II
MRR SCORES FOR EACH SINGLE LEARNER, LENGTH-ORIENTED CLASSIFER AND BOTH ENSEMBLES (CONFIGURATION III)

better than a syntax-oriented ensemble.

For length-oriented groups, the Bayes Classifier got the best performance for queries composed of 1, 2 and 5+ terms, whereas MaxEnt did well for queries containing 2 and 3 terms. For each group, Borda Counts outperformed the corresponding single learner, and MaxEnt Ensemble outperformed Borda Counts. Note that the increase of MRR scores is proportional to the number of terms, suggesting that the more context is provided by the query, the higher the performance gets. By looking at the number of query expansion features selected for each length (these were chosen 23 times by length-2 single learners, whereas these were chosen 10-11 times by single classifiers targeting longer queries), the lack of context is observed to radically affect two-term queries, and consequently, query expansion features become key items to resolve the intention for this type of query.

Experiments also showed that the most significant query expansion features included $Wikipedia$ sense discriminators, and the WEX categories, indicating useful query expansion features: few terms, likely one term, that signal the topic of the query.

	Borda Counts	MaxEnt Ensemble
Configuration I	0.9128	0.9143
Configuration II	0.9124	0.9156
Configuration III	0.9163	**0.9211**

TABLE III
OVERVIEW OF PERFORMANCE FOR THE PROPOSED ENSEMBLES

Single classifers aimed at two-term queries, relying on a large number of class features, as they provide useful context and narrow coverage. For instance, WEX categories were only found for 8.57% of the elements of this group. As for one-term queries, they are 93.70% of the times navigational, which makes it easier to guess their intent, whereas two-term queries are 37.61% and 56.50% informational and navigational queries, respectively.

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

For query intentions, the SVM got the best results when dealing with informational intentions (0.9521); Bayes did well for navigational (0.9063), and MLP did well for resources (0.8630). Overall, features extracted from dependency trees and NER tasks were significant for building effective intention classifiers. It is worth noting that three out of the four single classifers used the pair *partial relations* and NER domains, when resolving the queries' intention containing 4 and 5+ terms. Furthermore, WEX first-level categories and brand names were significant to three out of the four single classifers dealing with queries composed of two terms.

It was very difficult to infer intentions for web queries using short-length queries made of two terms, specially NP queries (only 39.78% of the NP queries had two terms, and 76.03% of two-term queries are NP). Thus, the MRR score for MaxEnt Ensemble for this intersection was only 0.8525, whereas it was 0.8647 for the remaining 23.97%.

Thus, our results for intent classification indicate that length-based ensembles are the best choice as the classifier selector only needs the token count to select the right set of classifiers, while at the same time it achieves the best overall performance.

V. Conclusions

This work has proposed a new multi-class ensembling strategy for automatically recognizing a user's intentions behind web queries. Our approach combines stochastic machine learning techniques and two ensemble methods in order to take advantages of multiple features extracted from different sources including knowledge bases, the query and other electronically available resources.

Experiments using our model assess different configurations for features, ensembling methods and classifiers showing that combining classifiers' outcomes assists in improving the quality of the user's intentions measured as position in a ranking of the best candidate intentions. Configurations of ensembles were composed of targeted classifiers, i.e., single classifiers aimed at specific lengths and syntactic patterns, indicating that designing ensembles with focused classifiers improved the ranking of user's intentions as compared with single classifier approaches.

Incorporating a 'classifier selection' task performed very well when comparing with other classification methods which may be due to that the feature optimization algorithm is capable of filtering those elements that are more suitable to each region. Even though some features might be included in several specialized classifiers, the distribution of its values might radically differ from one region to the other. As a natural consequence, focused classifiers can capture these differences across intentions more effectively. In real-life applications, it is a key factor to tailor search results that fits the display in small modern devices such as tablets and mobile phones.

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

ACKNOWLEDGEMENTS

This research was partially supported by FONDECYT (Chile) research project no. 11130094: *"Bridging the Gap between Askers and Answers in Community Question Answering Services"* and granted to Alejandro Figueroa, and FONDECYT (Chile) research project no. 1130035: *"An Evolutionary Computation Approach to Natural-Language Chunking for Biological Text Mining Applications"* granted to John Atkinson.

REFERENCES

[1] Areej Alasiry, Mark Levene, and Alexandra Poulovassilis. Extraction and evaluation of candidate named entities in search engine queries. In *WISE*, pages 483–496, 2012.

[2] Cory Barr, Rosie Jones, and Moira Regelson. The Linguistic Structure of English Web-Search Queries. In *Empirical Methods in Natural Language Processing*, pages 1021–1030, 2008.

[3] Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.*, 25(2), April 2007.

[4] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002.

[5] Junwu Du, Zhimin Zhang, Jun Yan, Yan Cui, and Zheng Chen. Using search session context for named entity recognition in query. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, 2010.

[6] Alejandro Figueroa and Guenter Neumann. Exploiting user search sessions for the semantic categorization of question-like informational search queries. In *International Joint Conference on Natural-Language Processing*, pages 902–906, 2013.

[7] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 267, New York, New York, USA, 2009. ACM Press.

[8] I. Hernández, P. Gupta, P. Rosso, and M. Rocha. A Simple Model for Classifying Web Queries by User Intent. In *2nd Spanish Conf. on Information Retrieval, CERI-2012*, pages 235–240, 2012.

[9] Bernard J. Jansen and Danielle L. Booth. Classifying Web Queries by Topic and User Intent. In *CHI*, pages 4285–4289, 2010.

[10] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2008.

[11] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 64–71, New York, NY, USA, 2003. ACM.

[12] K Krammer and Y Singer. On the algorithmic implementation of multi-class svms. *Proc. of JMLR*, 2001.

[13] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004.

[14] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 391–400, New York, NY, USA, 2005. ACM.

[15] Ying Li, Zijian Zheng, and Honghua Kathy Dai. Kdd cup-2005 report: Facing a great challenge. *ACM SIGKDD Explorations Newsletter*, 7(2):91–99, 2005.

[16] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Q 2 c@ ust: our winning solution to query classification in kddcup 2005. *ACM SIGKDD Explorations Newsletter*, 7(2):100–110, 2005.

[17] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24(3):320–352, July 2006.

[18] Wang Ting-Xuan and Lu Wen-Hsiang. Identifying popular search goals behind search queries to improve web search ranking. In *Proceedings of the 7th Asia conference on Information Retrieval Technology*, AIRS'11, pages 250–262, 2011.

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE

Alejandro Figueroa is a researcher at Yahoo! Labs, Santiago, Chile. He is also a research associate at Diego Portales University, Chile. He received his Ph.D in Computational Linguistics from Universitaet des Saarlandes, Saarbruecken, Germany (2010). Dr. Figueroa has also been a researcher at DFKI (the German Center for Artificial Intelligence), Saarbruecken, Germany, and Yahoo! Research Lab in Barcelona, Spain. His research interests include Question-Answering Systems, Natural Language Processing, Machine Learning and Information Retrieval.

John Atkinson is a full Professor of the Department of Computer Sciences, Universidad de Concepcion, Concepcion, Chile, and received his PhD in Artificial Intelligence from University of Edinburgh, Scotland, UK. He is actively involved in basic and applied research in Text Mining, Natural Language Processing, Artificial Intelligence, and Machine Learning. Dr. Atkinson is a member of the AAAI, the IEEE Computer Society, and senior member of the ACM.

Digital Object Indentifier 10.1109/MIC.2015.22 1089-7801/$26.00 2015 IEEE