Predicting Contextually Appropriate Venues in Location-Based Social Networks

Jarana Manotumruksa[1], Craig Macdonald[2], and Iadh Ounis[2]

School of Computing Science, University of Glasgow, G12 8QQ, UK
[1]j.manotumruksa.1@research.gla.ac.uk,
[2]{firstname.lastname}@glasgow.ac.uk

Abstract. The effective suggestion of venues that are appropriate for a user to visit is a challenging problem, as the appropriateness of a venue can depend on particular contextual *aspects*, such as the duration of the user's visit, or the composition of the user's travelling group (e.g. alone, with friends, or with family). This paper proposes a supervised approach that predicts appropriateness of venues to particular contextual aspects, by leveraging user-generated data in Location-Based Social Networks (LBSNs) such as Foursquare. Our approach learns a binary classifier for each dimension of three considered contextual aspects. A set of discriminative features are extracted from the comments, photos and website of venues. Using a dataset from the TREC 2015 Contextual Suggestion track, supplemented by venue annotations generated by crowdsourcing, we conduct a comprehensive experimental study to identify the set of features appropriate for our problem and to evaluate the effectiveness of our proposed approach. Our results demonstrate both the accuracy of our classification approach in predicting suitable contextual aspects for a venue, and its effectiveness at making better venue recommendations than the best performing system in TREC 2015.

1 Introduction

Making effective venue recommendations that a user may wish to visit relies on contextual information about the user, such as the user's location, time of visit, and previous venues visited. Dey *et al.* [7] defines context as "*any information that can be used to characterize the situation of an entity that is considered relevant to the interaction between a user and an application*". In the *context-aware venue recommendation* (CAVR) task, the involved entity is the user, whose context can be explicitly provided by the user or implicitly detected by sensing devices (e.g. GPS location). Moreover, CAVR is a challenging task, as users may not have visited a city before, rendering collaborative filtering approaches less useful. Therefore, to suggest venues to the users, approaches for effective personalised CAVR can encompass venue features (e.g. the number of people visiting the venue (check-ins) in an LBSN), user features (e.g. the user's rating of similar venues) and contextual features (e.g. the user's location and the time of the day).

In this paper, we argue that by considering new aspects of context, e.g. the duration of trip, the season of the year and the group of people the users are intending to visit the venue with, we can improve the effectiveness of a personalised CAVR system. However, unlike information about a venue's category or the number of check-ins, which are easy to obtain from LBSNs, identifying the appropriateness of venues to various contextual dimensions may not be directly made from the existing metadata of the venue in the LBSNs. We propose a personalised CAVR system that can account for contextual preferences explicitly provided by users, and which operates in two phases: firstly, leveraging user-generated data from a LBSN to predict appropriate contextual dimensions for each venue, using a supervised approach; and secondly adapting a state-of-the-art venue recommendation system to account for each venue's predicted dimensions when ranking venue suggestions. Moreover, as a venue can be appropriate for multiple dimensions of a contextual aspect, e.g. a restaurant is suitable to visit at day time and night time, this problem can be addressed as a *multi-label classification* problem. Indeed, we develop classifiers for the dimensions of three contextual aspects used in the recent TREC 2015 Contextual Suggestions track: (1) *Duration*, how long a trip the user is on? (2) *Season*, when is the most suitable season the user should visit the venue? and (3) *Group*, who is the venue suitable to visit with (e.g. with family)? In particular, to the best of our knowledge, the prediction of contextual dimensions for the Group aspect for a venue is a new problem that has not been addressed in previous works. Later, we show how to effectively integrate the proposed dimension classifiers as features within a CAVR system based upon learning-to-rank. In tackling this problem, this paper's contributions are as follows: (1) a learned approach that can predict appropriate contextual dimensions for a venue, based on different types of features, namely *temporal features* extracted from the venue's comments and photos on the LBSN, as well as *term-based features* extracted from the comments about the venue and the textual contents of the venue's website; (2) a demonstration of the usefulness of taking contextual aspects into account during venue ranking, based upon a TREC 2015 dataset. Indeed, the experimental results demonstrate the accuracy of our classification approach in predicting suitable contextual aspects for a venue and its effectiveness at making better venue recommendations than the best performing systems participating in TREC 2015.

2 Related work

Various existing works have shown that leveraging user-generated data in LBSNs can significantly enhance the effectiveness of context-aware venue recommendation (CAVR) systems (e.g. [5, 6, 14]). Yuan *et al.* [14] developed a collaborative *time-aware venue recommendation* that suggests venues to users at a specific time of the day. In particular, they mined historical check-ins of users in LBSNs to enable personalised venue recommendations using a time-aware collaborative filtering approach. Deveaud *et al.* [6] made time-aware venue recommendations by forecasting the popularity of nearby venues in the immediate future. However, all these approaches only considered the user's location and the time of the day as context when making venue recommendations. Recently, Hashem *et al.* [9] proposed an approach that recommends a sequence of venues to visit to users, which aims to optimise recommendation quality based on constraints (i.e. number of people, travelling time and distances). In contrast, we propose an approach that applies a learning to rank technique to recommend venues to users by considering multiple contextual aspects such as duration of the trip and type of the group the user like to travel with, rather than the number of users who are joining the trip.

Previous works on CAVR [13, 14] used check-in data from LBSNs to evaluate the effectiveness of their recommendation systems, by assuming that users implicitly like the venues they visited. However such data may not be appropriate to evaluate CAVR systems because check-in data do not explicitly express the users' contextual preferences. Indeed, research into CAVR has been boosted by the TREC Contextual Suggestion track [4]. This track aims to investigate search techniques for complex information needs that are highly dependent on the users' contexts and interests. In particular, the task addressed by the track is as follows: given the user's preferences (ratings of venues) and context (user's location), produce a ranked list of venue suggestions for each user-context pair. Moreover, in TREC 2015 [4], new contextual aspects were proposed. Additional contextual dimensions are provided by each user for each aspect: namely the duration and season of their trip, the type of trip (holiday, business etc.) and type of group the user is travelling with. Our work directly proposes an accurate modelling of the appropriateness of venues w.r.t. the aspects proposed in TREC 2015.

A few participants in TREC 2015 attempted to explicitly model the contextual appropriateness of the venues. Indeed, as the best performing participant, Aliannejadi *et al.* [2] proposed a system that learns the user's positive and negative profiles for the venues in the user's preferences, based on the positive and negative comments and categories defined by different LBSNs of the venues. However, they do not explicitly model the user's preferences in terms of aspects of contextual preference. McCreadie *et al.* [10] is the most similar to our own work in that they also examine the timestamps of photos and comments from an LBSN, but without using such evidence to predict the appropriate dimensions of context for a venue. In contrast, we propose to predict the contextual appropriateness of a venue (Section 3), by leveraging the photos and comments about the venue, as well as the content of the venue's website (Section 4). We later show how this can be used in making better context-aware venue recommendations (Section 6).

3 Problem statement

We now define the problem of predicting the appropriate contextual dimensions for a venue. Firstly, let V be a set of venues $\{v_1, \ldots, v_n\}$ and A be a set of contextual aspects about which users may express explicit requirements for relevant venue suggestions. In this work, we focus upon three contextual aspects proposed within the TREC 2015 Contextual Suggestions track [4], namely the *Duration* and *Season* of the user's visit, and the *Group* that the user intends to visit the venue with. Associated with each contextual aspect $a \in A$ is a set of dimensions, $a = \{d_{a,1} \ldots d_{a,m}\}$. Table 1 describes the dimensions for each of the contextual aspects. Therefore, the problem of predicting the appropriate contextual dimensions for a venue can be defined as follows: for a given venue v_i, predict the members of the set D_i, where D_i is the set of all contextual dimensions that the venue is appropriate for, i.e. $D_i = \{d | d \in a,\ \forall a \in A\}$. Indeed, each venue may be appropriate to multiple dimensions for a given contextual aspect, e.g. for the Season and Duration aspects, a park might be suitable to visit in the Spring or Summer, and only during the day time. We assume that each dimension is independent, e.g. a bar can be open both during the day and at night. Therefore, we formulate our problem as a *multi-label classification* problem and apply the most widely-used method by considering the prediction of each dimension as an independent binary classification problem, i.e. for a venue v_i, each $d \in D_i$ is identified by a binary classifier $h_d : v_i \rightarrow \{d, \neg d\}$.

4 Contextual Aspect Features

In this section, we describe our proposed approach that predicts the dimensions of contextual aspects that are appropriate for each venue. Our approach is based upon the

Aspect	Dimension	Description
Duration	Day Time	Is a venue suitable to visit between 6:00 AM - 6:00 PM?
	Night Time	Is a venue suitable to visit between 6:00 PM - 6:00 AM?
	Weekend	Is a venue suitable to visit on weekend?
Season	Spring	Is a venue suitable to visit between March and May?
	Summer	Is a venue suitable to visit between June and August?
	Autumn	Is a venue suitable to visit between September and November?
	Winter	Is a venue suitable to visit between December and February?
Group	Alone	Is a venue suitable to visit alone?
	Friends	Is a venue suitable to visit with friends?
	Family	Is a venue suitable to visit with family?

Table 1. The 10 dimensions of the contextual aspects that we consider in this work.

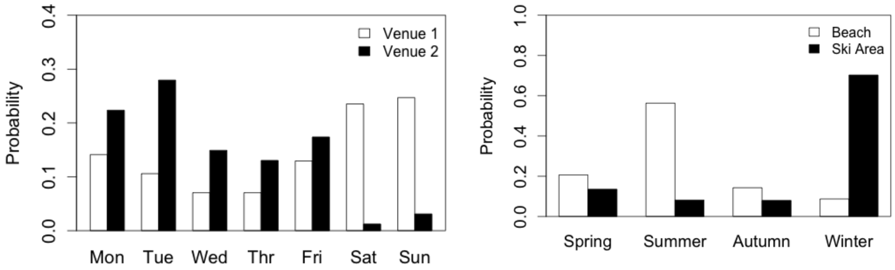

Fig. 1. Distribution of timestamps over different time patterns

definition and extraction of categorical and temporal features (Section 4.1) as well as textual features (Section 4.2) that are suitable for training the 10 binary classifiers, i.e. one for each dimension of the contextual aspects in Table 1.

4.1 Categorical and Temporal Features

Intuitively, due to the different activities offered by each venue, different venues generally exhibit different temporal characteristics e.g. a venue such as a bar is more suitable to be visited at night time, while a venue such as a park is more suitable to visit during the day. Such intuitions can be used to extract *temporal features* for each venue. In LBSNs, users can upload photos taken at a venue they are visiting or write a comment to review the venue they have visited. The timestamps of comments (photos) and the venue's metadata (e.g. venue's categories) can be leveraged to extract discriminative features for each venue, which will be used to train our binary classifiers.

In terms of notation, each venue v_i has a set of associated comments $R_i = \{r_1 \ldots r_n\}$, and photos $P_i = \{p_1 \ldots p_m\}$, as well as a set of categories $\Theta_i = \{\tau_1 \ldots \tau_n\}$ and a website W_{v_i}. Both a photo p_j and a comment r_k are represented as a tuple $\langle u, v, t, content\rangle$, indicating that the photo or comment is generated by user u at venue v at time t, where $content$ represents the actual image of the photo or the text of the comment. The time t (e.g. "2015-02-15 15:45:22") that either the photo or comment was generated is represented as a *time-slot*, for instance as a specific hour of the day (15:00), a day of the week (Sunday) or a month of the year (February). $TS_m(t)$ is a function that returns *time-slot* w.r.t. the specific *time-slot* granularity m, e.g. this function can be chosen to produce a time slot for each hour of the day, i.e. $TS_{hour}(t) \in \{0, 1, \ldots, 23\}$. From now on, the term timestamps is equally applicable to the timestamps of comments or photo, unless otherwise specified. Next, we propose to extract *category features* and *temporal features* for the Duration and Season aspects, based on the venue's metadata and the timestamps of photos or comments uploaded by LBSN users.

Category Features ($f1$, $f2$): Intuitively, venues belonging to a similar category likely share similar contextual appropriateness to each other. $f1$ is a feature indicating the category membership of a venue within the 10 top-level Foursquare categories[1]. Similarly $f2$ represents the membership of the 147 low-level Foursquare categories.

Temporal Venue-based Feature ($f3$): The timing of visits by users to venues differs and can be indicative of its appropriateness to different contexts, e.g. a venue mostly visited at weekend is less likely to be appropriate for a weekday. Figure 1(a) provides an example of the distribution of timestamps of 2 venues over the days of the week, demonstrating that the venues exhibit different temporal characteristics. Hence, for a given venue v_i, we calculate the maximum likelihood probability of observing comments (or photos) with a timestamp that is appropriate for a dimension d of a time-based aspect (Duration or Season), $p(v_i|d) = \frac{\sum_{j \in R_i}^{n} AT(TS_d(j), d)}{|R_i|}$, where R_i (P_i) is the set of comments (photos) for venue v_i, and $AT(.)$ is a function that returns 1 if timestamp j is appropriate to a given contextual dimension d, 0 otherwise, based on the time descriptions listed in Table 1.

Temporal Category-based Features ($f4$,$f5$): $f3$ suffers from a sparsity problem – as most venues in our dataset have a small number of comments/photos in the LBSN – thereby hindering the accuracy of a classifier using this feature. To alleviate this problem, we assume that similar venues share similar contextual behaviour, e.g. all ski park venues are more likely to be appropriate to visit in winter rather than in summer, while all beaches are more suitable to visit in summer (this can be seen in Figure 1(b)). In particular, we calculate the likelihood at the level of a category τ, $P(c|\lambda) = \frac{\sum_{v_i \in V} P(v_i|\tau) \cdot P(v_i|\lambda)}{\sum_{v_i \in V} P(v_i|\tau)}$, where $P(v_i|\tau)$ is a binary function denoting if venue v_i belongs to the given category τ (1 if true, 0 otherwise). Note that we consider as separate features the distribution of top-level ($f4$) and low-level ($f5$) Foursquare categories.

4.2 Term-based Features

Unlike the *temporal features* described above, we cannot use timestamps to infer the appropriateness of a venue for dimensions of the Group aspect. In this section, we describe our term-based features for the Group aspect that score occurrences of appropriate terms within two sources of evidence, the websites and the comments of venues.

Web-based Term Feature ($f6$): Intuitively, if a venue wishes to attract a particular audience, its website will contains terms related to the corresponding dimension(s) of the Group aspect. For instance, a restaurant website that contains "family deals" in its menu section is likely to be appropriate to visit with family. To illustrate this, Figure 2 shows how terms relating to each dimension of the Group aspect occur within two venues that we have identified as suitable for Family and Friends respectively. Indeed, from the figure, it can be seen that the website of a venue suitable for a family group exhibits a higher frequency of terms relating to that dimension than a venue suitable for

[1] https://developer.foursquare.com/categorytree

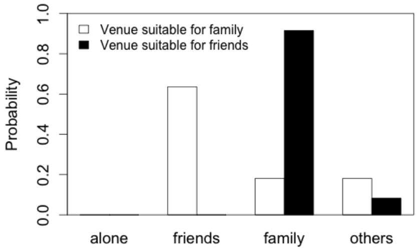

Fig. 2. The distribution of term frequency of two venues on the Group aspect.

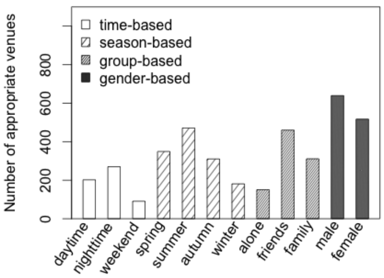

Fig. 3. The distribution of appropriate venues for each dimension of contextual aspects in crowdsourcing dataset.

friends does, and vice versa. Therefore, the occurrence of terms corresponding to each dimension of the Group aspect in a venue's website is likely to be a useful feature for predicting the appropriate Group dimensions of venues. To extract discriminative features for the dimensions of the Group aspect, we collect terms related to each dimension from an external web resource[2]. We then index the websites of venues (extracted from the venue metadata using a standard IR system, and issue to the system a query Q_d consisting of a set of terms corresponding to the dimension of context d. Finally, we use the system's retrieval score of each venue's website for each dimension of context as a single feature, $P_{term}(v_i|d) \propto score(Q_d, W_{v_i})$, where Q_d is a query consisting of the set of terms related to the given dimension d of the Group aspect, W_{v_i} is the website of venue v_i and $score(.)$ is a standard retrieval model. Hence, the higher score the more likely the venue is suitable for the dimension of contextual aspect.

Comment-based Term Features ($f7$, $f8$): These features are defined similarly to $f6$, except that the comments for each venue are indexed, instead of the venue's website. However, users may vary in the sentiments they express in their comments about venues they have visited. Ignoring these sentiments may hinder the classification performance. For instance, a venue that contains a negative comment like "*I was disappointed that there were no small chairs for children*" will obtain a high retrieval score since its comment contains family-related terms, although this venue is likely not appropriate to the Family dimension. To tackle this limitation, we use the SentiStrength [12] sentiment analysis tool, which was developed for short user-generated content such as tweets, to classify all of the comments of the venues into three different classes: positive, negative, and neutral. We then separately index the positive and negative comments for each venue, while ignoring the neutral comments. Features $f7$ and $f8$ are calculated as for $f6$, but for the the positive and negative comments, respectively. Next, we evaluate the accuracy of our proposed contextual dimension classifiers (Section 5). Later, in Section 6, we show that learned ranking approach with contextual features generated from our proposed classifiers can significantly outperform the best TREC participants.

5 Venue Dimension Classification Evaluation

In this section, we evaluate the accuracy of the classifiers through answering two research questions: **(RQ1)** Can we exploit the distribution of timestamps of photos or

[2] http://www.enchantedlearning.com/wordlist/

comments to predict appropriate contextual aspects for venues for the Duration and Season aspects? **(RQ2)** Can we leverage the terms occurring in either the venue's website or comments to predict the appropriate dimensions of the Group aspect for venues?

5.1 Crowdsourcing Venue Annotations

We use crowdsourcing to obtain ground truth data by asking workers to annotate the dimensions of context suitable to venues. We randomly select 746 venues from the TREC Contextual Suggestion 2015 test collection. We use the CrowdFlower[3] crowdsourcing platform, asking workers to annotate the applicable contextual dimensions for each venue, based upon representative information of each venue extracted from the Foursquare LBSN. In particular, for each venue, the worker views the venue's title, category, an image and two randomly-selected comments, and uses check-boxes to indicate appropriateness for each contextual dimension. Following best practices for crowdsourcing [1], and to ensure the quality of the obtained ground truth data, we ask three different workers to label each venue, resulting in 2,238 judgements, for a total cost of US$31[4]. The distribution of judgements for each dimension is shown in Figure 3. The final annotations are derived by choosing the dimensions of context that the maximum number of workers agreed upon, e.g. if 2 workers agree that a venue is suitable to visit in Spring and Summer while 1 worker considers that the venue is suitable to visit in Winter, the final ground truth dimensions for that venue are Spring and Summer.

5.2 Experimental Setup

Learning Algorithms. We use the Weka machine learning software [8] for training and predicting contextual dimensions. We explore the effectiveness of our classifiers using 3 classification algorithms: Naive Bayes, J48, and SVM. All classification experiments are conducted using a 10-fold cross-validation on the crowdsourcing dataset.
Retrieval Models. To extract the *term-based features* $f6 - f8$, we index the venues' websites and comments using v4.0 of the Terrier platform[5] and use BM25 for calculating $score(.,.)$. While other standard weighting models can be used, initial experiments found that our conclusions are not changed by the choice of weighting models.
Evaluation Measure & Baseline. We report the accuracy of our contextual dimension classifiers for each dimension in terms of the F_1 classification measure. As the problem of contextual dimension classification has never been addressed before, and as the nature of our dataset is imbalanced across the class labels of each dimension, we compare our proposed approach with a baseline that classifies each venue as the majority class for each dimension (denoted as Majority), i.e. for all dimensions except weekend, the majority class would be 'appropriate' (see Figure 3).

5.3 Experimental Results

Firstly, Table 2 reports the accuracy, in term of F_1, of contextual dimension classification using different classification algorithms learned with all features across the Duration, Season and Group aspects. For brevity, we report mean F_1 across all dimensions

[3] http://crowdflower.com [4] Our crowdsourced venue annotations are freely available from http://dx.doi.org/10.5525/gla.researchdata.325 [5] http://terrier.org/

F_1	Duration	Season	Group	Mean	Δ
Majority	0.481	0.488	0.541	0.503	
Naive Bayes	**0.680**	**0.573**	**0.574**	**0.609**	
J48	0.602	0.542	0.548	0.564	-7.88%**
SVM	0.482	0.489	0.542	0.504	-11.97%**

Table 2. F_1 accuracy of contextual dimension classification using different classification algorithms. Δ differences denoted by * exhibit significant decreases (McNemar's test, $p < 0.01$) compared to Naive Bayes.

F_1		Duration			Season					
		day time	night time	weekend	spring	summer	autumn	winter	Mean	Δ
Majority		0.342	0.465	0.638	0.382	0.642	0.342	0.588	0.486	
All	comments	0.628	0.689	**0.723**	**0.532**	**0.656**	**0.563**	**0.604**	**0.627**	
	photos	**0.644**	**0.695**	0.72	0.502	0.538	0.532	0.525	0.595	
Ablation										
-$f1$	comments	0.615	0.691	0.717	0.521	0.652	0.530	0.598	0.618	-1.59%
	photos	0.635	0.700	0.725	0.496	0.536	0.531	0.517	0.591	-0.670%**
-$f2$	comments	0.598	0.676	0.711	0.514	0.649	0.530	0.612	0.613	-2.39%**
	photos	0.634	0.699	0.722	0.462	0.493	0.518	0.451	0.568	-4.54%**
-$f3$	comments	0.611	0.696	0.720	0.527	0.664	0.549	0.609	0.625	-0.48%
	photos	0.653	0.687	0.732	0.494	0.513	0.509	0.486	0.582	-2.18%**
-$f4$	comments	0.645	0.672	0.720	0.541	0.649	0.491	0.61	0.618	-1.59%
	photos	0.630	0.687	0.724	0.519	0.661	0.547	0.529	0.614	3.19%*
-$f5$	comments	0.619	0.690	0.701	0.505	0.643	0.578	0.604	0.620	-1.27%
	photos	0.644	0.693	0.729	0.504	0.649	0.571	0.638	0.633	6.39%**

Table 3. Classification accuracy in terms of F_1, for each source of evidence (comments or photos), for each dimension of the Duration and Season contextual aspects. Majority denotes classification using only the majority class, while All denotes all features. Feature groups from All are ablated. Best performances for each dimension are highlighted in bold. Values denoted by * and ** exhibit significant differences (McNemar's test where $p < 0.05$ and 0.01 resp.) compared to All features for each source of evidence.

for a given aspect. Recall that our Majority baseline is where all instances in the test set for a given dimension are classified as the majority class. For this table, we use a default experimental setting, which we vary below: we use timestamps from the comments to extract the time-based features ($f4$ & $f5$). Indeed, Table 2 shows that Naive Bayes significantly outperforms both J48 and SVM in predicting the appropriate dimensions for venues across all aspects. The high effectiveness of Naive Bayes when trained with a small dataset is also supported by the literature, e.g. [3]. Hence, in the remainder of our analysis and experiments, we focus solely upon the Naive Bayes classifier.

Next, Table 3 reports the classification accuracy in terms of F_1 for each dimension of the Duration and Season aspects, for each source of evidence (comments or photos). The top part of the table reports effectiveness when using only the Majority class (our baseline) as well as when all features ($f1$-$f5$) are used for these aspects (denoted All). On analysing this part of the table, we note firstly that the F_1 scores are markedly higher than the Majority class baseline. Moreover, while effectiveness is slightly higher when using the timestamps of photos for the Duration aspect, the timestamps of comments are overall more effective, providing more valuable evidence for the Season aspect. This

Features	alone	friends	family	Mean	Δ
Majority	0.576	0.667	0.382	0.542	
All	0.600	0.564	0.557	0.574	
$f1$ & $f2$	0.644	0.700	**0.594**	0.65	11.20%*
+ $f6$	0.661	**0.709**	0.590	**0.653**	12.19%*
+ $f7$	**0.671**	0.695	0.575	0.647	11.33%*
+ $f8$	0.668	0.673	0.580	0.640	10.41%*

Table 4. For the Group aspect, the table reports F_1 for different feature combinations, as well as Δ w.r.t. F_1 score of All, where * exhibit significant increase (McNemar's test, $p < 0.01$).

fits with our intuition of the mobile-phone based use of the Foursquare LBSN: users are likely to upload photos when they are currently attending the venue. In contrast, comments are often left after the user has visited the venue, perhaps reflecting on a good or bad time he/she had at the venue. This makes the timestamps of comments less useful for accurately predicting the appropriate dimensions of the Duration aspect.

The second part of Table 3, denoted Ablation, reports F_1 when groups of features are ablated (removed) from All features, with the column Δ reporting the mean increase or decrease compared to the corresponding classifier using All features. In general, the largest decreases in effectiveness are obtained when the low-level category information $f2$ is removed from the features used by the dimension classifiers, showing that detailed knowledge of the venue category can be informative in accurately predicting the appropriate dimensions for venues. Features $f3$ (for photos) and $f4 - f5$ (for comments) are also shown to be important, but comparatively less so.

For the Group aspect, Table 4 follows a similar layout to Table 3. In the top part of the table, we report the F_1 classification effectiveness for All applicable features for this aspect (namely $f1$, $f2$, $f6$, $f7$, $f8$). Recall that $f6 - f8$ are *term-based features*, extracted from venue's website, positive comments and negative comments, respectively. In Table 4, we observe that this contextual aspect represents a more difficult classification problem, where the majority class is comparatively strong (Mean F_1 0.542 over the three dimensions). The results show that, on average, our classifiers are more effective than the majority in predicting the appropriateness of venues for the group aspect.

Next, the second part of Table 4 reports different combinations of features, starting with the categorical features $f1$ & $f2$ alone, and then adding $f6 - f8$ (each calculated using BM25) in turn. We observe that the F_1 scores for the combination of features are overall higher than for the All features, suggesting that more data would be required to obtain the most effective model. Moreover, among the *term-based features* $f6 - f8$, $f7$, which is calculated using the positive comments offers the highest improvement over $f1$ & $f2$ for the alone dimension. For friend dimension, the textual contents of the website, $f6$, offers the largest margin of improvement. Overall, the general trend across all rows in the bottom part of Table 4 is that the textual evidence from the websites is more important than the positive comments ($f7$), which is in turn more important than the negative comments ($f8$). This surprising result can be explained in that the comments are sparse in comparison with venues' websites. Indeed, the number of tokens indexed from websites and comments index are 17,138,495 and 1,515,640, respectively.

To summarise, our findings for research question **RQ1** were that the *temporal features* - based on the timestamps of the comments and photos for each venue - can be useful for creating accurate classifiers for the Duration and Season aspects (as shown in Table 3). For **RQ2**, we find that textual evidence found on the websites of the venues is the most useful on average for predicting the appropriate dimensions of the Group aspect.

6 Ranking Contextually Appropriate Venues

In this section, we describe how we improve the effectiveness of a CAVR system using our contextual dimension classifiers trained on *temporal features* extracted from the timestamps of comments and *term-based features* extracted from venue websites.

Firstly, we formally describe the venue recommendation scenario of the TREC 2015 Contextual Suggestion track [4] in which our evaluation is conducted. Rankings in the Contextual Suggestion track are created in response to a user-context pair, denoted $\langle U_j, C_j \rangle$ (and which can be thought as a "query"). A user's profile consists of a set of venue preferences, denoted as $U_j = \{v_i \rightarrow p_{i,j},\ \ldots\}$, where $p_{i,j}$ is the user's preference rating (1 to 5) for venue v_i. The context C_j contains a number of contextual preferences in terms of the dimensions of interest to this work: $C_j = \{d\}$. As only one dimension can be expressed for each of the aspects listed in Table 1, $|C_j| = |A| \leq 3$.

Given a set of dimensions preferences C_j (e.g. $\{Weekend, Summer, Alone\}$) expressed by the user, we now describe how we integrate the outcome of our dimension classifiers into the ranking approach of a CAVR system. Firstly, following recent work in creating personalised venue suggestions [5], we adopt a learning to rank approach to take into account different sources of evidence when ranking venues, by making use of features about the venue, $\mathcal{F}(v_i)$ and features representing how the venue matches the users interests, $\mathcal{F}(v_i, U_j)$, (e.g. a cosine similarity between the categories of the venue v_i and the categories of the venues rated positively in U_j). Moreover, we encompass the expressed contextual preferences as one numerical feature for each aspect, denoting the confidence of classifier that the venue is appropriate for dimension $d \in C_j$:

$$\mathcal{F}(v_i, C_j) = \bigcup_{a \in A} \begin{cases} h_d(v_i) & \text{if } d \in C_j \wedge d \in a, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $h_d(v_i)$ is the *confidence* of the classifier for dimension d that venue v_i is appropriate for d. Note that not all user-context pairs express a contextual dimension for each aspect. Hence, when no dimensions of contextual aspect a are present in C_j, the classifier confidence for that aspect a is replaced by 0, therefore $\forall C_j$, $|\mathcal{F}(v_i, C_j)| = 3$.

6.1 Experimental Setup

In the following, we address a final research question: (**RQ3**) Can our proposed dimension classifiers improve the effectiveness of a state-of-the-art CAVR system? Our experiments make use of the TREC 2015 Contextual Suggestion track test collection. As our venue ranking features rely on information about the venue from Foursquare, we only consider venues in the TREC test collection that originated from Foursquare. Our baselines are the two best approaches from TREC 2015, mentioned in Section 2 (namely uogTr [10], USI [2]). For a fair comparison, we also remove venues suggested by these TREC participants that did not originate from Foursquare, as well as any users in the test collection who did not explicitly express any contextual preferences (i.e. $|C_j| = 0$). This results in 194 user-context pairs in the collection (down from 211 pairs).

As a basis for our experiments, we use a personalised CAVR system based upon learning to rank – similar to that of Deveaud *et al.* [5] – building upon the Automatic Feature Selection (AFS) [11] technique that creates a linear learned model. This model is trained on the 60 venue preferences of all users – as these are separate from the test venues, this represents a clear separation between training and test environments. We report the effectiveness using the measures reported in the track overview paper [4], namely Precision@5 (P@5) and mean reciprocal rank (MRR). For each venue, we calculate a total of 11 venue ranking features, namely 6 venue features ($\mathcal{F}(v_i)$): number of check-ins, number of likes, number of comments, number of photos, average Foursquare rating, unique number of users - all obtained from the Foursquare API, 2 user-venue features ($\mathcal{F}(v_i, U_j)$): cosine similarity between the categories of the venue v_i and the categories of the venues rated by the user in his/her profile U_j – one feature for positive-rated venues, and one for negative-rated venues, and 3 contextual features ($\mathcal{F}(v_i, C_j)$): Classifier confidences for the dimensions expressed by the user in C_j.

6.2 Experimental Results

Table 5 indicates the sources of evidence considered by each of the systems in terms of user-venue preferences (denoted as User), venue information (Venue), and contextual sources of evidence (Context). The first part of Table 5 shows the effectiveness of the learned model obtained from AFS using all 11 venue ranking features (denoted All), as well as when different feature groups have been ablated. In this table, we observe that the best overall results are achieved by the model trained with All features. Moreover, ablating the contextual features generated by our 10 dimension classifiers causes a decrease in P@5 (-2.45%), showing the importance of these features in an effective ranking model. We also observe the same significant decrease in effectiveness (upto 11%) when venue features are removed, also reported by Deveaud *et al.* [5].

	User	Venue	Context	P@5	Δ	MRR	Δ
TREC Median	-	-	-	0.5090		0.6716	
AFS (All)	✓	✓	✓	**0.6020**		**0.7858**	
AFS (VC)	×	✓	✓	0.6010	-0.17%	0.7827	-0.40%
AFS (UC)	✓	×	✓	0.5443*	-10.60%	0.7394*	-6.28%
AFS (UV)	✓	✓	×	0.5876	-2.45%	0.7800	-0.74%
uogTr	✓	✓	✓	0.5742*	4.84 %	0.7584	3.61%
USI	✓	✓	×	0.5722	5.21%	0.7494	4.86%

Table 5. The effectiveness of learned CAVR using different features, in comparison with the 2 best TREC 2015 Contextual Suggestion track systems. The performances denoted * exhibit significant decreases in effectiveness (paired t-test, $p < 0.05$) compared to the All feature.

Next, we compare the effectiveness of the learned CAVR models with the two best performing systems in the TREC 2015 Contextual Suggestion track. We find that the AFS model trained with All features outperforms the best TREC 2015 participants, for both P@5 and MRR. Note that without our proposed contextual features (AFS (UV)), our CAVR system would have not significantly outperformed uogTr approach. Indeed, while the uogTr is similar to ours, it does not deployed learned classifiers for identifying contextual appropriateness of venues. Hence, for **RQ3**, we find that our proposed classifiers can markedly enhance an CAVR system and can significantly outperform the best participating TREC 2015 system in terms of P@5 and MRR.

7 Conclusions

In this paper, we proposed classifiers that can predict the appropriateness of venues to contextual dimensions, and showed how they could be successfully integrated into a state-of-the-art CAVR system. Our results showed not only that dimensions can be accurately predicted, but that by considering the new dimensions of context, the quality of venue recommendation can be significantly enhanced. Moreover, we found that textual contents of venue's website is more suitable than comments about the venue on an LBSN for identifying if the venue is suitable to visit under different dimensions of the Group aspect, while the temporal characteristics of venues can be successfully captured using the timestamps of comments or photos. A direction for future research will encapsulate the modelling of dependencies between aspects of contextual dimensions.

References

1. Aker, A., El-Haj, M., Albakour, M.D., Kruschwitz, U.: Assessing crowdsourcing quality through objective tasks. In: Proc. of LREC. ELRA (2012)
2. Aliannejadi, M., Bahrainian, S.A., Giachanou, A., Crestani, F.: Univ of Lugano at TREC 2015: Contextual suggestion and temporal summarization tracks. In: Proc. of TREC (2015)
3. Brain, D., Webb, G.: The need for low bias algorithms in classification learning from large data sets. In: Proc. of PKDD (2002)
4. Dean-Hall, A., Kamps, J., Kiseleva, J., Voorhees, E.: Overview of the TREC 2015 contextual suggestion track. In: Proc. of TREC (2015)
5. Deveaud, R., Albakour, M.D., Macdonald, C., Ounis, I.: On the importance of venue-dependent features for learning to rank contextual suggestions. In: Proc. of CIKM (2014)
6. Deveaud, R., Albakour, M.D., Macdonald, C., Ounis, I.: Experiments with a venue-centric model for personalised and time-aware venue suggestion. In: Proc. of CIKM (2015)
7. Dey, A.K., Abowd, G.D., Salber, D.: A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-computer interaction (2001)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
9. Hashem, T., Barua, S., Ali, M.E., Kulik, L., Tanin, E.: Efficient computation of trips with friends and families. In: Proc. of CIKM (2015)
10. McCreadie, R., Vargas, S., Macdonald, C., Ounis, I., Mackie, S., Manotumruksa, J., McDonald, G.: Univ of Glasgow at TREC 2015: Experiments with Terrier in contextual suggestion, temporal summarisation and dynamic domain tracks. In: Proc. of TREC (2015)
11. Metzler, D.A.: Automatic feature selection in the markov random field model for information retrieval. In: Proc. of CIKM (2007)
12. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. JASIS&T 63(1) (2012)
13. Yao, L., Sheng, Q.Z., Qin, Y., Wang, X., Shemshadi, A., He, Q.: Context-aware point-of-interest recommendation using tensor factorization with social regularization. In: Proc. of SIGIR (2015)
14. Yuan, Q., Cong, G., Ma, Z., Sun, A., Thalmann, N.M.: Time-aware point-of-interest recommendation. In: Proc. of SIGIR (2013)