Keyword based categorisation of diary entries to support personal Internet content pre-caching on mobile devices

Andreas Komninos
University of Strathclyde
26 Richmond Street
Glasgow G1 1XH UK
+44 141 548 3160
andreas@cis.strath.ac.uk

Mark D. Dunlop
University of Strathclyde
26 Richmond Street
Glasgow G1 1XH UK
+44 141 548 3160
mark.dunlop@cis.strath.ac.uk

ABSTRACT

This paper presents a study into the effectiveness of our algorithm for automatic categorisation of real users' diary entries, as a first step towards personal Internet content pre-caching on mobile devices. The study reports an experiment comparing trial subjects allocations of 99 diary entries to those predicted by a keyword-based algorithm. While leaving considerable grounds for improvement, results are positive and show pave the way for supporting mobile services based on categorising users' diary entries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models, Query formulation*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Calendars, keywords, query formulation, personal pre-caching

1. INTRODUCTION

Early research [2],[3],[1] has indicated that the entries found in calendars, tend to fall under specific categories. To be able to decide on the nature of the information held in a calendar entry, it is important to identify the keywords contained therein; this would allow assumptions to be made regarding the category that the entry belongs to and thus services that may be helpful to the user could be provided on a category –by-category basis, for example in our work to pre-cache mobile devices with internet content that may be useful to the user.

There has been research carried out in the area of enabling intelligent agents for mobile devices to make web searches and present them to the user when offline[10],[11]. However In previous work [1] we reported a model that would allow the pre-caching of Internet content on a mobile device, without requiring any explicit instruction from the user. The implementation of this model results in a predictive system, which relies on the information found in a user's calendar, in order to make assumptions for their daily activities and the internet content that might be needed to support them. The model described in [1] attempts to categorise calendar entries and then formulate relevant searches and retrieve documents that are relevant to the category, as defined below. These searches are formed through the combination of keywords extracted from the entry and search terms that are relevant to a category. In our system, each category comprises of a "descriptor", a list of keywords that are clues that an entry might belong to a category, and also a list of "search terms", keywords that are common in web queries that are relevant to the category the entry belongs to. In order to illustrate the system's operation, consider an entry that contains the keyword "Buenos Aires" in the location field, a city that is different from the user's known base location. The system may assume that the user is travelling there and combine "Buenos Aires" with keywords such as "map", "hotel" or "flights", which are known to be common searches on the Internet for travel destinations.

In this paper, an experiment is described that was conducted to test the capability of our categorisation algorithm to correctly identify diary entry categories and to investigate the way real people make assumptions on the categorisation of calendar entries. The disparities between machine and human results are analysed and useful information is provided for the refinement and ideal operation of the machine agent.

2. CATEGORY ASSIGNMENT

In order to develop our proposed model [1] we collected more than 200 entries from real users' calendars/diaries, both electronic and paper-based. All of the entries in these calendars were provided at random by the users in our presence. We also informed the users that we would be keeping their entries confidential and anonymous. Hence, we can reasonably assume that there was no deliberate effort from the users to hide sensitive entries. In fact, some users even went as far as to export and provide us with whole unedited sections of their electronic calendars.

These calendars belong to 20 members of staff and students at the University of Strathclyde. The entries were analysed and,

Mobile and Ubiquitous Information Access Workshop, September 2004, Glasgow, UK

with the help of the information provided by the people who wrote the entries, were assigned to 9 categories: Birthdays, Class (to attend), General task (to do), Meeting (group & personal), Miscellaneous, Reminder, Social, Travel, and Work-related task[4]. To alleviate any confusion as to category names, Table 1 shows the nature of entries encompassed by each category.

It is important to stress here that these categories reflect the opinions of the calendar users, who are highly familiar with the context of their entries. One can observe that there is some overlap between the calendar entries, for example, Birthday is a subset of Social. However, where such overlap is maintained, it is because there was a strong indication from the users that such a low-level category is significant and should exist separately from its high-level parent. Further to these findings, we created a categorisation system based on keyword extraction from diary entries to automatically classify diary entries into these 9 categories, as described in section 3.

3. AUTOMATIC CATEGORISATION OF CALENDAR ENTRIES

Based on the analysis of calendar entries from real users, as described in section 2, we have determined the existence of 9 categories of entry types. Each of our pre-determined categories has been assigned a list of identifying keywords and key phrases, based on the lectical analysis of the contents of our calendar entries. Our lists of keywords include items such as verbs, nouns, common names, surnames and also rules for the combination of these items. The keywords come from the manual examination of user entries. We call these lists the "category descriptors". This grouping of relevant items is inspired from the clustering methods in information retrieval[6],[7].

Further more, each category has also been assigned a list of terms, which have been identified as relevant and common searches on the internet for items that may belong to a category. During the collection of calendar entries, we asked the users what kind of Internet search they might have performed on the given entry. The analysis of their answers provided a list of search terms that are appropriate for each category. These search terms are also complemented by terms that have been obtained from Google's keyword suggestion tool[5], especially for the traveling category.

During the analysis of a calendar entry, the algorithm examines each word individually and assesses the following information:

- Whether the word belongs to a category descriptor.
- The location of the word in the entry (e.g. whether it's in the notes, title or location field).
- Whether the word is followed or preceded by certain other words, therefore forming a phrase that belongs to a category descriptor

This information is weighed in relevance to its importance. According to the answers to these findings, each category is assigned a score, indicating the probability that the entry belongs to that given category. At the end of the analysis of the entry, these scores are compared and item is identified as belonging to the category that has the greatest score. The algorithm then associates the identified keyword with several search terms, in order to formulate web queries and fetch relevant documents.

The process of adapting a query to better suit search engines (query re-formulation) is not a new one [8],[9]. Here, we solve the problem of adapting a query to a user's needs and optimizing it for a general search engine by obtaining the query context through inference from the categorization of the calendar entry.

Table 1 : The categories and their description

Category	***Description***
Birthday	Indicates someone's birthday
Class (to attend)	User has to attend a class (either as a student or lecturer)
General Task (to-do)	General tasks to complete (non-work related), such as buy an item or email someone
Meeting (group & personal)	A meeting that has to be attended or an appointment with someone
Miscellaneous	Unclassifiable items
Reminder	A reminder that an event is happening, such as "Mary is off sick" or "Exams start today"
Social	A social event, such as dinner or going to the movies
Travel	User has to travel to some destination out of their habitual location
Work – related task	A task to do that is related to the user's work, such as "write a report" or "mark exam scripts"

4. TEST EXECUTABLE AND DATA SET

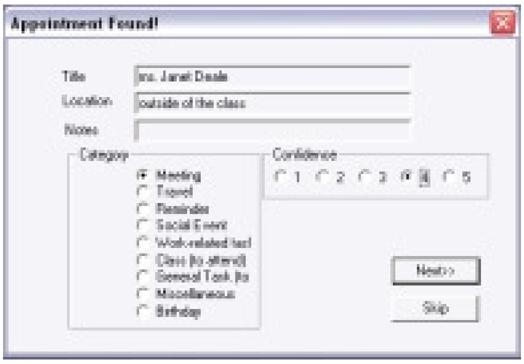
To assess the quality of this automatic categorisation we wrote a small application that would read in a pre-compiled static list of entries and present them to the user one at a time. For each entry, the user would be asked to assign a category to each of the entries and to provide the level of confidence that accompanied their decision (figure 1). To limit study time and user frustration the collection was sub-sampled by approximately 50%, giving a total number of entries of 100 per subject. The entries provided in the test collection were randomly selected from the original collection while preserving the category distribution (see table 2). In the original collection some entries had two or more instances caused by recurring appointments – these were preserved in the test collection; we chose not to eliminate any duplicate entries, in order to observe the perseverance of the test subjects to their original perception of the category recurring entries belong to.

Below we present sample screenshots of the actual test executable. Figure 1 shows the entry categorisation dialog, with the entry details displayed on the top part. The user is obliged to select one choice from the Category radio button group and one from the Confidence group.

Table 2: The entry categories and their numerical representation in the test collection

Category	%	*Category*	%
Birthdays	4	Reminder	11
Class (to attend)	4	Social	9
General task (to do)	6	Travel	8
Meeting (group & personal)	48	Work-related task	6
Miscellaneous	4		

Figure 2 shows the dialog displayed when further classification information is asked from the user. The entry details are displayed again, along with the user's choices from the previous dialog. The user can enter further information in the box at the bottom of the dialog.



Appointment Found!

Title: ms. Janet Deale

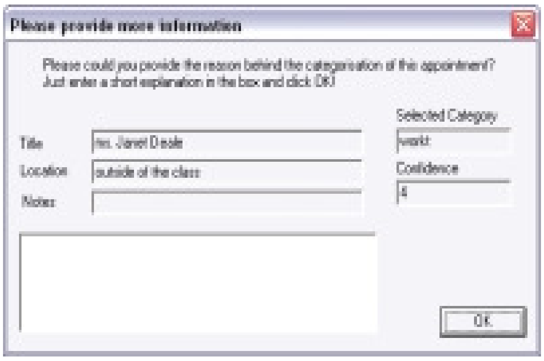
Location: outside of the class

Notes:

Category
● Meeting
○ Travel
○ Reminder
○ Social Event
○ Work-related task
○ Class (to attend)
○ General Task (to
○ Miscellaneous
○ Birthday

Confidence
○ 1 ○ 2 ○ 3 ● 4 ○ 5

Figure 1: Entry categorisation screen



Please provide more information

Please could you provide the reason behind the categorisation of this appointment?
Just enter a short explanation in the box and click OK!

Title: ms. Janet Deale

Location: outside of the class

Notes:

Selected Category: work

Confidence: 4

Figure 2: Explanation screen

When the user clicks on Next in Figure 1, the application compares the category allocation with the keyword-based prediction of the entry's category. If the user's choice agrees with the prediction, then the next entry is displayed. Otherwise, the user is asked to provide some more information on the rationale behind their choice (Figure 2). To reduce users from feeling that they were somehow "wrong" and to encourage them to feel confident and honest in the explanation provided, they were told these explanations would be requested randomly. The information from each assessment of is logged using an XML data structure.

5. REALISATION OF THE EXPERIMENT AND DATA PROCESSING

5.1 Test subjects

The test program was distributed to 10 individuals to run unsupervised on their own computers. The test subject group did not include any of the original providers of appointment entries in order to eliminate any possibility that some of the ambiguous entries were familiar to the subjects. This was done in order to simulate the algorithm's natural uncertainly and unfamiliarity with the user's environment. The test subject group consisted of 2 postgraduate students, 2 non-academic professionals and 6 undergraduate students. Finally, the algorithm was allowed to run a simulation of the user interaction on the test collection, independently, and produce its own log of results, which would be compared against the correct choices and the choices of the subject group.

5.2 Notes

It should be noted here that there really isn't an objectively "correct" choice for most of the entries, since the meaning of an entry strongly depends on the user's context. However, by saying "correct", we mean to describe the categorisation of entries, as provided by their originators.

Another important note regarding the selection of the test subjects is that we were not trying to assess whether the system performs adequately for the users that provided the original entries. Such an attempt would be extremely biased due to the fact that the original system is based on the information gathered from these people. It could be therefore expected that the system would perform in a satisfactory manner, providing however results that are meaningless due to the bias. The aim of the pre-caching system is to begin with some initial knowledge, which will, in time, adapt to a user's particularities through the process of implicit relevance feedback. The scope of this experiment is to test the sufficiency of the basic knowledge for initial system operation in unfamiliar user environments and contexts.

Unfortunately, due to an error in the execution of the test program on one subject's computer, one of his entries was lost. Therefore, to maintain a consistent result, that entry was excluded from analysis, resulting in 99 entries tested per subject – as the entry was of type "Meeting", there was little impact on the frequency distribution of categories from that given in table 2.

6. RESULTS AND ANALYSIS

6.1 Analysis targets

The post-processed data collected for each calendar entry is displayed in the following graphs (figure 3):

- The number of answers that disagreed with the correct choice
- The number of answers that disagreed with the algorithm's choice
- Whether the most popular choice agrees with the correct choice
- Whether the most popular choice agrees with the algorithm's choice

This data is reported for all levels of confidence and separately for all the answers that were made with a confidence level greater or equal to 4. This was done in order to reduce the impact of lucky guessing. Also, it would be interesting to observe the level of confidence with which the users decide on their perception of the entries.

6.2 Original results

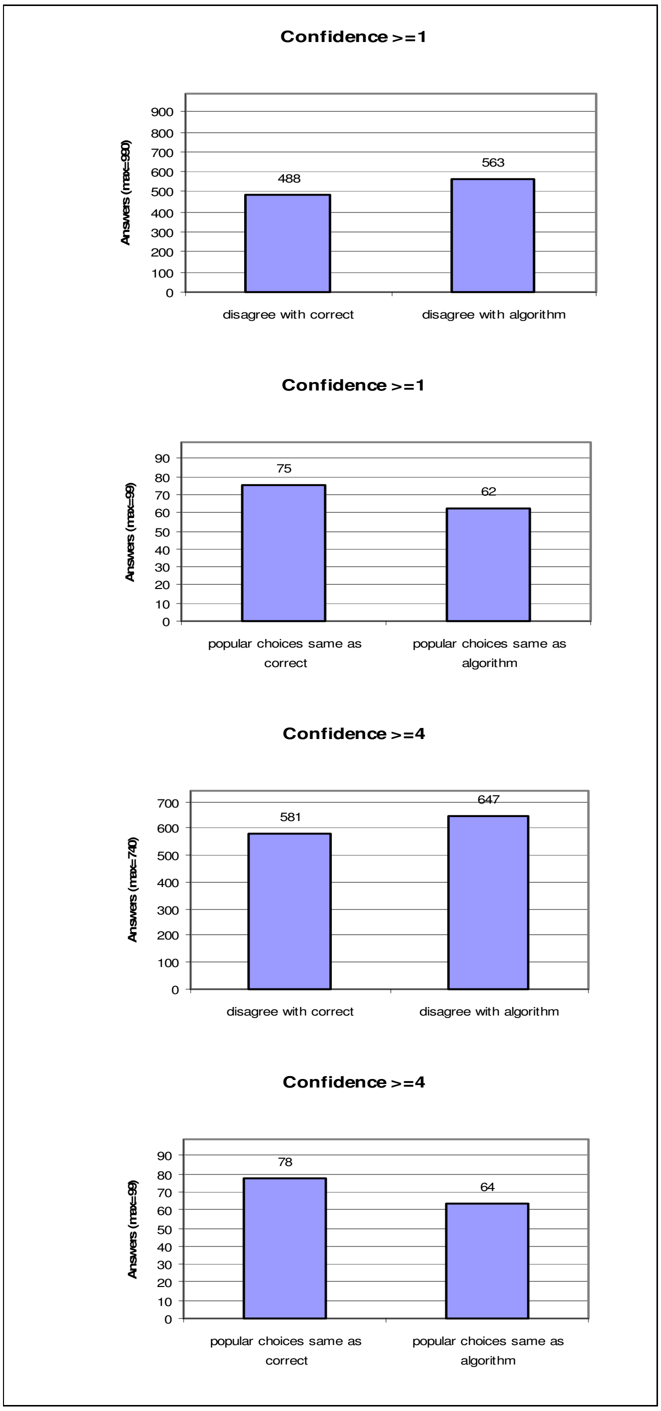


Figure 3: Comparison of categories with original users' allocations

Table 3: Summary of original results

	Confidence >=1	*Confidence >=4*	*Difference*
Agree with Correct	49.3%	44.7%	4.6%
Agree with Algorithm	56.9%	53.6%	3.3%
Popular choice same as Correct	75.8%	78.8%	3.0%
Popular choice same as Algorithm	62.6%	64.6%	2.0%
Algorithm's correct guesses	73.7%	n/a	n/a

6.3 Analysis of original results

Table 3 summarises these results. It is interesting to observe that less than half (49.6%) of the answers given by subjects actually coincide with the answers provided by the entry owners, a percentage which drops further to 44.7% when considering only the answers given with a strong degree of confidence. This could be interpreted as indicating that the success rate of our algorithm, without any training, should not have to exceed 50% to be considered equivalent to human performance.

Having measured the performance of our algorithm, on the same collection of test data, we have found it to score an approximate 73.7%, which is well beyond what the expectations based on the human performance should be. It is true that our algorithm has been formulated using, amongst other things, rules and keywords derived by the analysis of our original entry collection. However, we have found similar performance levels when executed on smaller collections of entries, which were obtained after the analysis of our initial collection.

Another interesting note is that while only half of the answers given actually were in accordance with the "correct" answers, when considering the most popular category choice for each entry, the percentage amounts to 75.8%. The close proximity of this number to the success rate of the algorithm shows that the algorithm is very close to electing the same choice as the "majority" of the subjects, therefore it is close to adopting the best/most appropriate elements of human rationale for the completion of its task.

It is noteworthy also to observe that the number of answers that were given with a high degree of confidence is rather large and makes up for approximately 75% (740) of the total amount of answers (990). From this we can conclude that users appear to be quite confident about their choices, even though less than half of them are "correct". Since the active prediction of the user's choice is paramount, in order to provide personalised and meaningful results that are particular to the user, it appears that the target for an acceptable, perhaps tolerable, success rate for the prediction and suggestion of categories, and thus related information, should lie in these levels of 75%.

Finally, for the 11 recurring appointments, we measured the most popular choices at each occurrence. It was discovered that with a confidence level greater than zero, for four of the appointments (36%) the category was changed but only one change was actually to a "correct" choice. With a confidence level greater or equal to four, three entries (27%) were given a different popular choice, with none coinciding with a "correct" choice. Naturally this sample is fairly small and one cannot be conclusive, however, this seems an unlikely large percentage.

6.4 Revised experiment design and results

In section 2, we mention the fact that there is some overlap between the categories as assigned by the entry owners, which could result in some ambiguity. The results, as described above, compare the low-level categorisation of the entries by their owners with that of people who are completely unfamiliar with the context under which the entries were made. It can therefore be argued that there might exist a bias towards error, against the test subjects. To remove such a suspicion, we analysed the same results again, having grouped the categories Class, Work Task and General Task under a more general category called TASK, and also the categories Social and Birthday, under SOCIAL. The revised results can now be summarised as follows:

Table 4: Revised summary results

Revised results	*Confidence >=1*	*Confidence >=4*	*Difference*
Agree with Correct	51.7%	42.2%	9.5%
Agree with Algorithm	59.2%	50.1%	9.1%
Popular choice same as Correct	74.7%	78.8%	4.1%
Popular choice same as Algorithm	62.6%	63.6%	1.0%
Algorithm's correct guesses	73.7%	n/a	n/a

Table 5 shows the comparison of the original versus the revised results. From this table, it is apparent that any differences, where they occur, are not significant as their magnitude is of approximately 2.5%. It can be argued therefore that the analysis of the original results is valid for the revised experiment, despite the change in the experiment design.

Table 5 : Comparison of Original and Revised Results

	Conidence. >=1		*Confidence >=4*	
	Original	*Revised*	*Original*	*Revised*
Agree with Correct	49.3%	51.7%	44.7%	42.2%
Agree with Algorithm	56.9%	59.2%	53.6%	50.1%
Popular choice same as Correct	75.8%	74.7%	78.8%	78.8%
Popular choice same as Algorithm	62.6%	62.6%	64.6%	63.6%
Algorithm's correct guesses	73.7%	73.7%	n/a	n/a

Furthermore, it can therefore be concluded that the breakdown of high-level categories into subcategories and that any overlap that may appear in the original categorisation scheme does not have any significant impact in the performance of the average test subject (our algorithm seems to remain thoroughly unaffected). However, the assignment of sub-categories might actually be desirable, as it would allow the production of more appropriate web searches. Indeed, our system is already designed to support inheritance from "master" or other categories.

7. CONCLUSIONS

This paper reports work on using a keyword-based algorithm for predicting the category of diary entries in order to support mobile services (e.g. pre-caching of probably relevant internet content). In particular the paper reports the results of an experiment comparing the effectiveness of the keyword-based algorithm with that of people other than the entry authors. The results show that, on average, people can individually correctly classify a diary entry only 49% of the time, while the majority decision from a group of users achieves 75% accuracy when compared with the original entry's author's categorisation. Furthermore, our keyword-based algorithm achieved 73% accuracy – nearly matching that of the majority of users and well exceeding that for individual users.

The study report here was based on real diary entries from 20 people. While there is clearly ground for improving our algorithms, which we are investigating (e.g. use of long-term relevance feedback), we believe that correctly identifying categories at this level of performance will give a significant lead to mobile systems that are attempting to provide information and services based on the user's activity. Further work needs to address the effectiveness of our query formulation strategy and the assessment of the document relevance that it will produce. Our next study will investigate the performance of pre-caching internet content based on category information about diary entries (thus reducing the need for direct internet accesses while travelling).

8. REFERENCES

[1]KOMNINOS, A., DUNLOP, M.D. (2003): Towards a model for an Internet content pre-caching agent for small computing devices, *Proceedings of the 10th International conference on Human Computer Interaction (HCII2003), Crete, Greece, 2003*

[2]KELLEY, J.F, CHAPANIS, A (1982): How professional persons keep their calendars: Implications for Computerisation, *Journal of Occupational Psychology, 55, pp. 241-256*

[3]KINCAID, C. M., DUPONT, P.D., KAYE, A. R (1985): Electronic calendars in the office: An assessment of user needs and current technology, *ACM Transactions on Office Information Systems, 3(1), pp. 89-102*

[4]KOMNINOS, A., DUNLOP, M.D. (2004): Enhancing the usability of calendar applications in mobile communication devices through entry categorisation, *http:www.cis.strath.ac.uk/~andreas/paper2_5.pdf*

[5]GOOGLE KEYWORD SUGGESTION TOOL: https://adwords.google.com/select/main?cmd=KeywordSandbox (valid 1/6/2004)

[6] HAYES, R.M:, 'Mathematical models in information retrieval', in Natural Language and the Computer (Edited by P.L. Garvin), *McGraw-Hill, New York, 287 (1963).*

[7]VAN RIJSBERGEN, C. J.: Information Retrieval, *2nd edition, Dept. of Computer Science, University of Glasgow (1979).*

[8]BUDJIK, J., HAMMOND, K., Watson: Anticipating and contextualising information needs, *62nd Annual meeting of the American Society for Information Science, Medford, NJ, 1999.*

[9]MITRA, M., SINGHAL, A., Buckley, C.: Improving automatic query expansion, *ACM SIGIR 98, Melbourne, Australia, 1998.*

[10]MYERS, B.: Mobile Devices for Control, *4th International Symposium on Mobile HCI, Pisa, 2002*

[11]ARIDOR, Y., CARMEL, D., MAAREK, Y., SOFFER, A., LEMPEL, R.: Knowledge Encapsulation for Focused Search from Pervasive Devices, *ACM Transactions on Information Systems 20(1), p.p. 25-46*