List, Group or Menu: Organizing Suggestions in Autocompletion Interfaces

Alia Amin[1], Michiel Hildebrand[1], Jacco van Ossenbruggen[1],
Vanessa Evers[2], Lynda Hardman[1,2]

[1]Semantic Media Interfaces, CWI, Amsterdam, The Netherlands,
[2] University of Amsterdam, The Netherlands

Abstract. Even though Autocompletion is widely used in search interfaces with different variations, few guidelines exist on how to present autocompletion suggestions. In this paper, we describe two user studies that shows some types of organization strategies help users search faster and easier in a known-item search task. We studied the effect of grouping suggestions in an autocompletion interface to select terms taken from a thesaurus. In the first study, we explored ways of grouping term suggestions from two different thesauri (TGN and WordNet). The results of the first study suggested that the best organization strategies are different when using different thesaurus. Users found *Group* organization may be appropriate to organize location names from TGN, while *Alphabetical* works better for WordNet. We then carried out a follow-up study, where we compared three different organization strategies (*Alphabetical*, *Group* and *Composite*) for location name search tasks. The results of the second study indicate that in general, autocompletion interfaces help improve the quality of keyword. We also found that *Group* and *Composite* organization help users search faster, and users perceive the suggestion organization as easier to understand and use than the *Alphabetical*.

1 Introduction

There is a lot of interest in the Information Retrieval community for interactive query expansion features that can help improve user search performance and the quality of queries submitted. There are two types of interactive query expansion: real-time query expansion (provide suggestion during query formulation) and post-query expansion (provide suggestion after query formulation). Between the two types, real-time query expansion (RTQE), such as autocompletion has been most adopted in many operational search applications e.g. *Google Suggest* or *Yahoo! Search Assist*. RTQE is an attractive feature because it can reduce the required number of keystrokes, decreases the user's cognitive load through term recognition (rather than recall) and helps the user avoid typing errors [7, 9]. Additionally, it can improve the quality of initial query for known-item as well as exploratory tasks [17, 18]. RTQE is better then post-query expansion because it lowers task completion time, increases search engagement and increases uptake of interactive query expansion [17]. Most research efforts are directed towards improving query expansion suggestions, e.g.[1, 5, 13, 19], and generally paid less attention on the interface issues. Many RTQE stick to only *list* organization as presentation style. Prior works have lead us to believe that different types of implementation of RTQE presentation would likely result in different user search performance. In [4] three different interfaces on top of the same retrieval system were compared. The study suggested that the quality and effectiveness of search depend on the combination between both the retrieval system and its interface in supporting query expansion. Joho et al. [10] studied different query expansion presentation styles. They compared two types of organization strategies: *list* of alphabetically ordered and *menu hierarchy* interactive query expansion interfaces. They found that even though there is no significant difference in the precision-recall between using the two interfaces, people finish the search task significantly faster when using the *menu hierarchy*. Another study [11] compared two different hierarchical IQE systems (based on the subsumtion approach and trigger phrased on parent-child description) against a baseline (no suggestion). They found that accessing the hierarchies reduces search iterations, reduces paging actions and increases the chance to find relevant items. In practice, many variants of RTQE organization strategies have been deployed, such as:

- **List** organization strategy, such as by *alphabetical* list (WikiSearch), by *popularity* (popular query/destination in [16])
- **Group** organization strategy, such as Google Suggest uses 2 groups: personal history query and popular prefix match, or `Naver.com` uses 2 groups: popular prefix and suffix match.
- **Menu** organization strategy, such as a cascaded menu hierarchy in [10, 11]

In this research, we focus on the presentation aspect of an autocompletion, namely organization strategies and how they influence users search performance. We are motivated by the usage of relationships of terms from a thesaurus to improve RTQE presentation. Certain relationships between terms from a thesaurus has been known to improve the quality of query expansion. Efthimiadis et al. [5] investigated the terms used in a IQE for the INSPEC database. They reported that variants (synonym) and alternative terms (i.e. narrower, boarder and related terms) relationships are useful for query expansion. Similarly, Joho et al. [10] in their research found that for WordNet, the most useful relationships are hyponym, hypernym and synonym. Additionally, they also mention conceptual relation (e.g. teeth-dentist) as a meaningful relationship for query expansion. In this study, we investigate how to improve selection of terms in autocompletion interfaces. In particular, we explore the potential of hierarchical relations in thesauri to improve the organization of autocompletion suggestions. By imposing grouping and ordering strategies we provide a means of navigating the suggestions and finding the right terms faster and easier. We restrict our discussion to interfaces that syntactically complete the input based on exact or partial literal match. We do not consider *query recommendation* which tries to extrapolate queries based on certain (semantic) relations or algorithms, such as in [3]. Semantic relations are used, if at all, primarily to organize the suggestions in the interface. We carry out two related studies. The first examines the quality of grouping strategies for different thesauri, the second investigates to what extent grouping and (alphabetical) ordering are able to influence the suggestion selection process.

2 Organization of Suggestions

In this section, we discuss the look and feel of alternatives organization strategies for autocompletion suggestions used in Study 1 and Study 2 (see Fig.1 and Fig. 2) [1]. Examples are taken for TGN [2] autocompletions, similar visualization and algorithm is applied to WordNet.

Alphabetical order — Fig.1a shows autocompletion suggestion in an alphabetical order. The suggestions are organized in the following priority: prefix match on primary literal (location name), prefix match on secondary literal (country name), e.g. the suggestion "*Paris*, Canada" is shown before "*Paris*, France". Exact match are presented first, followed by partial match. The first part of suggestion consists of 15 items. When the user selects the "see more" button, all suggestion is presented as a long list.

Group — An organization that combines similar suggestions under a common heading. The grouping category is conveyed visually as a group title. Where terms are related by explicit thesaurus relations, any of these relations can be used as a basis for grouping. Grouping can be based on variants of hyponym relations. There are 2 types of grouping: predefined and dynamic. In predefined grouping the category is always of the same type. For example, TGN's hierarchy is based on geographical containment (e.g. *Europe* > *France* > *Paris*). Grouping can be based on any predefined level within this hierarchy, e.g. grouping by country (Fig. 1b). Alternatively, predefined category can be based on common property, such as place type (Fig. 1c) e.g. inhabited place (city, village) or body of water (stream, lake).

Another variant is the dynamic grouping where the group heading differs and is determined by an algorithm that optimized groups based on the number of suggestions retrieved. The desired groups can be preset taken from the top level hierarchy (Dynamic 1) or taken from the lowest (leaf) level hierarchy (Dynamic 2). Dynamic groups could provide an alternative grouping for thesauri with irregular hierarchical structure such as WordNet. Fig. 1d is an example of Dynamic 1 group implemented for TGN.

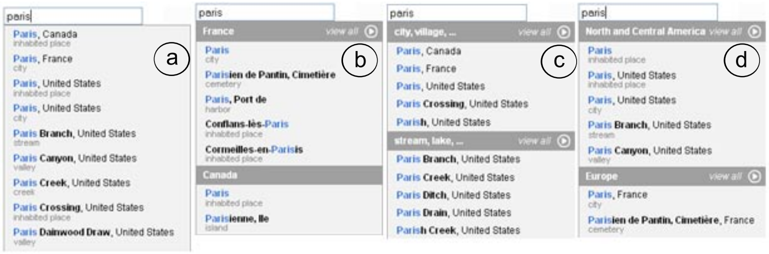


Fig. 1. Autocompletion with different grouping strategies used in Study 1 for TGN, from left to right: a) alphabetical order, b) by country, c) by place type, d) by continent

[1] Autocompletion demo is available at `http://anonymous.org/demos/`

[2] Thesaurus for Geographical Names `http://www.getty.edu/research/conducting_research/vocabularies/tgn/`

Composite — A composite organization resembles a two level cascaded menu hierarchy. It is groups similar suggestions into a single item (primary menu), deferring their display to a secondary menu. Fig. 2d shows an example composite suggestion interface, the primary menu contains all exact match of all location names from the same country. The secondary menu displays further information of the location names that allows disambiguation e.g. *Kingston (the city)* or *Kingston (the parish)*. This strategy retains the simplicity of alphabetical order, but shows larger numbers of alternatives in the limited amount of screen real estate available.

3 User Studies

We conducted two user studies to investigate the benefits and trade-offs of different strategies for organizing suggestions in autocompletion interfaces. The first study was an exploratory study to investigate the effects of grouping strategies on two different thesauri. The second study built on the result that grouping terms in a thesaurus of locations can be beneficial. The study investigated different organization strategies on the same set of suggestions.
Technology — The study was done using our autocompletion widget[3] that works on all major browsers supporting (X)HTML, CSS and Javascript. The client side widget is an extension of the Yahoo User Interface autocompletion widget (YUI v. 2.3.1[4]). The suggestion server has been implemented using SWI-Prolog's web infrastructure[5]. The autocompletion architecture is fully described in [2].

3.1 Study 1: Grouping Strategies

In Study 1, we investigate different variations of grouping as a type of suggestion organization. As mentioned in the previous section, there are many alternative implementations of grouping strategy using various term relationships in a thesaurus. The goal of Study 1 is to investigate to what extent grouping strategies for autocompletion suggestions can be applied to thesauri and if so, which grouping strategies are meaningful for users. We have chosen to implement similar grouping strategies for two different thesauri: a domain-specific thesaurus, TGN and a global thesaurus, WordNet. Our intention was not to compare the two thesauri, but to evaluate the suitability of different group strategy when implemented for these thesauri. Our research questions in Study 1 are: Can group organization strategy be implemented for the thesauri? Which group strategy is perceived the best by users?
Interfaces — We selected 4 autocompletion interfaces to compare with each other for TGN and similarly 4 for WordNet. The four chosen interfaces for each thesaurus are those which we thought were best to offer to users after informal trials of different algorithms and combinations. For TGN, the grouping strategies are: grouping by country (Fig. 1b), grouping by place type (Fig. 1c) and Dynamic 1 grouping (Fig. 1d). As a

[3] Demo is available at `http://anonymized.org/demo/`

[4] `http://developer.yahoo.com/yui/autocomplete/`

[5] `http://www.swi-prolog.org/`

baseline, we used the Alphabetical order of suggestions (Fig. 1a). We refer these interfaces as the Location Name (LN) interfaces.

For WordNet, the 3 grouping strategies are: predefined grouping using the top nine WordNet category nouns from the hypernym hierarchy, and two dynamic groupings: Dynamic 1 and Dynamic 2. Similarly, the Alphabetical order was chosen as a baseline. We refer these interfaces as the Object name (ON) interfaces.

Participants — Participants were recruited by sending out invitations to universities and research institutes from diverse departments, such as computer science, engineering and natural science. In total, 47 people responded. Participants were mostly students and some university employees. All participants reported that they use the Internet daily and are familiar with the autocompletion interfaces (e.g. in email clients, search engines and web browsers), 14 participants have experience with autocompletion interfaces in specialized applications such as script editors and interactive script interpreters.

Procedure and Tasks — The study was done as an online interactive experiment. All session activities are logged. Prior to the task, participants answered short questionnaire focusing about their experiences with autocompletion. Afterwards, every participant was assign tasks with 4 four TGN-LN interfaces (within subject design). For every LN, participants were asked the same tasks: to formulate several location queries, such as *Berlin* (city name) or *Alps* (mountain system)), and find the correct location names from the suggestions presented in the interface. Afterwards they were encouraged to try out their own example queries and explore the interface responses. After completing the tasks, participants were asked to answer assessment questions about the quality of the groupings and to give their comments. Finally, participants were asked to rank their preferred strategy for LN, from the most to the least preferred, and provide reasons for their decisions. The same task and procedure are repeated by the participants for the WordNet-ON interfaces. Participants were asked to formulate object queries, such as *Barbecue* or *Party*, and answer the assessment questions about the quality of grouping in this interface. The assessment questions on the quality of the group organization were derived from criteria taken from the literature [6–8, 12, 14]. The answer of the assessment questions are given in a 7-Likert scale (1:low, 7:high). These criteria are:

Q1 - perceived similarity of items within the same group; *"I think the items belonging to each group in this type of lists are similar to each other."*

Q2 - perceived difference of items between groups; *"I think the items belonging to different groups in this type of lists are different from each other."*

Q3 - affinity item and group title; *"I think the relationship between the items and group title is clear in this type of list."*

Q4 - reasonable number of groups; *"I think the number of groups in this type of list is appropriate."*

Q5 - group title appropriateness; *"I think the titles of the groups in this type of list are clear."*

The order of the interfaces were counter balanced using the Latin Square scheme among the participants. Pilot sessions were conducted to ensure that the participants could perform the tasks and understood the questions. The time to complete the study was approximately 30 minutes.

Results — The data we collected from the experiment were processed qualitatively and quantitatively. Our server log indicates that in addition to trying all provided examples, additionally some participants explored the behavior of autocompletion interfaces by trying out their own examples, such as different cities, countries or river names (e.g. Rhein) for LN and various object names (e.g. muscle, mobile, partner) for ON. It is important for us to confirm that participants explore the behavior of the autocompletion beyond the given task before assessing the quality of the interfaces.

Table 1. Left: Mean assessment scores, Right: Preferred grouping strategy (n=47 people, Study 1).

TGN (LN)	Mean Score (*SD*) *			
Question	Place type	Country	Dynamic 1	*p-value*
Q1	**5.30(1.68)**	4.57(1.83)	4.34(1.75)	.03
Q2	5.00(1.52)	4.53(1.80)	4.51(1.52)	.71
Q3	5.77(1.49)	5.74(1.51)	5.49(1.57)	.39
Q4	4.91(1.77)	**4.15(1.98)**	4.98(1.76)	.02
Q5	5.30(1.79)	**5.94(1.41)**	5.19(1.85)	.03
WordNet (ON)	Mean Score (*SD*) *			
Question	Predefined	Dynamic 1	Dynamic 2	*p-value*
Q1	4.19(1.56)	4.21(1.85)	3.94(1.65)	.77
Q2	4.64(1.47)	4.43(1.60)	**3.96(1.43)**	.01
Q3	4.13(1.81)	4.28(1.75)	4.13(1.66)	.61
Q4	4.19(1.72)	**3.47(1.73)**	4.02(1.88)	.01
Q5	3.83(1.81)	4.04(1.71)	3.72(1.82)	.48

* 7-Likert scale, score 1:strongly disagree, 7:strongly agree

TGN (LN)	Mean Rank (*SD*)	*p-value*
Place type	2.23(1.15)	.16
Dynamic 1	2.35(1.09)	
Country	2.67(1.13)	
Alphabetic	2.74(1.09)	
WordNet (ON)	Mean Rank (*SD*)	*p-value*
Alphabetic	**1.98(1.23)**	.02
Dynamic 1	2.62(.97)	
Predefined	2.68(1.09)	
Dynamic 2	2.72(1.06)	

• *Assessment:* The participants' assessments for six grouping strategies are shown in Table 1 (left). We examine each question to understand the characteristics of each grouping strategy using Friedman two-way analysis by ranks[6]. For LN we found that: (a) Place type grouping scored best with respect to perceived similarity - Q1 (χ^2(2)=7.36, *p*=.03)[7] (b) Country grouping scored best with respect to group title appropriateness - Q5 (χ^2(2)=6.77, *p*=.03)[8] (c) Country grouping scored lowest with respect to the number of groups - Q4 (χ^2(2)=8.11, *p*=.02) [9] Perceived similarity indicates the cohesiveness between the suggestions in a group. Place type grouping scores highest for this aspect. Alternatively, the Country group strategy gives most representative group titles (Q5) but poor on the number of group (Q4). One disadvantage of our implementation for the Country group strategy is that we did not make any limitation in the number of groups allowed. Because of this, the autocompletion list can potentially be very long. This is an adjustable parameter of the interface and not an inhereted characteristic of the thesaurus. The assessment score indicates that from the 3 types of LN grouping, Country and Place type are relatively good grouping strategies that each excel in different qualities.

[6] Nonparametric statistics is used as the data did not meet parametric assumptions

[7] Wilcoxon signed ranks (WSR) *post-hoc* test result for Q1: Place type scored sig. higher than Dynamic 1 ($p \ll .05$).

[8] WSR *post-hoc* test result for Q5: Country scored sig. higher than Dynamic 1 ($p \ll .05$) and Place type (*p*=.03)

[9] WSR *post-hoc* test result for Q4: Country scored sig. lower than Dynamic 1 (*p*=.02) and Place type (*p*=.01)

For the ON interfaces, we found that: (a) Dynamic 2 group scored lowest with respect to perceived difference - Q2 (χ^2(2)=10.17,p=.01)[10] (b) Dynamic 1 group scored lowest with respect to the number of groups - Q4 (χ^2(2)=9.66, p=.01)[11] The results showed that none of the ON group strategies excels from each other in the assessment score. We only found that the Dynamic 1 and Dynamic 2 groups perform the worst in Q2 and Q4. We think this is because the dynamic group strategies actually add participant's cognitive burden when they are trying to go through the suggestion list. No grouping strategy in ON that is assessed the best by our participants. The reason for this will be clear in the next results where we compared all group strategies against a baseline (Alphabetical order) and examine users preference.

• *Preference:* Table 1 (right) shows the Mean Rank of each grouping strategy for LN and ON. A low Mean Rank score indicates most preferred, and a high score is least preferred. Using the Friedman two-way analysis by ranks, we found that there is no strong preference in any of the location grouping strategies (χ^2(3)=5.14, p>.05). From the comments provided by the participants, we see that participants prefer different interfaces for different reasons. We conducted the same analysis for the four ON interfaces and found a different result. Participants strongly preferred the Alphabetical order to all other organization strategies (χ^2(3)=10.38, p=.02)[12] From the participants' comments, we understood that they found it difficult to understand the ON grouping strategies. This could explain the strong preference for Alphabetical order.

• *Comments:* Participants' comments gave us an explanation as to their assessment decisions and preferences. It seems that most decisions on chosing a LN interface is based on personal preference. *"By country seems more logical and pragmatic. Place type takes some getting used to but could work fine. Dynamic (grouping) gets confusing, Alphabetical (list is) not very clarifying."* [P6]. For the ON interfaces, participants oppinion are more uniform. The main comment was that many participants struggle with understanding ON grouping strategies. The baseline (Alphabetical order) seems to be the easiest to understand based on their past experience with finding terms in a dictionary. *"... I am more familiar with encyclopedic or dictionary structuring. The problem with such group autocompletion advice is that the adaptation process is quite time costly."* [P25].

Retrospective — The main goal of study 1 was to get a feel for how users perceive different grouping strategies. More precisely, we want to find out if and how the different structures of the thesauri used effect the user's perception, and whether the resulting groupings make sense at all.

Ideally, the best grouping strategy are the ones which scores highest on all five assessment scores (Table 1 left) and most preferred (Table 1 right). However, this is not the case. For LN, we found that one grouping strategy is better in some aspect while others in another aspect. For ON, we did not found any outstanding grouping strategy.

Thus, we concluded that grouping strategies may not be suitable for every type of thesaurus. For a domain-specific thesaurus, such as TGN we could find a sensible grouping strategy that people could understand relatively easily. In a global thesaurus such

[10] WSR *post-hoc* test result for Q2: Dynamic 2 scored sig. lower than Predefined (p=.01)

[11] WSR *post-hoc* test result for Q4: Dynamic 1 scored sig. lower than Predefined ($p \ll .05$) and Dynamic 2 (p=.03)

[12] WSR *post-hoc* test result for Mean Rank of preference: Alphabetical scored sig. lowest (i.e. strongly preferred) then Predefined (p=.02), Dynamic 1 (p=.04) and Dynamic 2 (p=.01).

as WordNet, however, the results are different. The users preference, assessment scores and participants' comments lead to the conclusion that for WordNet the group organization may not be the best strategy to use. In cases where the underlying thesaurus does not provide the information necessary for appropriate grouping, an Alphabetical order is the best option.

3.2 Study 2: Organization Strategies

Based on what we have learned in Study 1, we conducted a follow up study. We narrowed down the scope of study 2 by only investigating autocompletion for TGN. We decided not use WordNet since none of our group strategies for WordNet outperformed the baseline (Alphabetical). The goal of the second study is to compare three types of autocompletion: Alphabetical order, Group and Composite. We would like to investigate which interface helps users to search for terms from a thesaurus the fastest and easiest. To be able to come to this conclusion, we setup an experiment where users are required to use autocompletion for known-item search tasks. To evaluate speed, we measure performance in time to complete task (objective measurement). To measure ease-of-use, we took three subjective measurements: user assessments, preference and comments. Additionally, we analyze the quality of keywords provided in each condition.

Interfaces — In this study, we compared 4 different interfaces, namely: Alphabetical order (Fig.2b), Group (Fig.2c), Composite (Fig.2d) and a no autocompletion (NAC) interface (Fig.2a).

Participants — We recruited participants in the same manner as in the first study. In total, 41 people participated. Participants were aged between 16-66 years (*M*=30.90, *SD*=10.45). In general, participants use the Internet frequently (*M*=34.96, *SD*=19.51) (hours per week), and have medium to high familiarity with autocompletion interfaces[13].

Procedure — Each participant is assigned interfaces: NAC, Alphabetic, Group and Composite (within subject design). The order of the conditions were counter balanced using the Latin Square scheme among the participants. Pilot sessions were conducted to ensure that the participants could perform the tasks and understood the questions. The time to complete the study is approximately 25-30 minutes. In the experiment, participants started by answering general questions about their experience in using the Internet and autocompletion. Participants were then given a trial session to get accustomed to the interfaces. During the experiment, participants were given 24 tasks. In every task, time measurements were taken and participants were asked to assess the usability of the interface afterwards. We are interested in comparing the usability aspects of the different interfaces. After every interface, participants answered two questions (5-Likert scale):

Q1 - *"I find this interface easy to use."*

Q2 - *"I find the organization of the suggestions easy to understand.*"

[13] 1:low familiarity, 5:high familiarity; Autocompletion in search engines (*M*=3.40,*SD*=1.34), email client (*M*=4.12,*SD*=1.17), address browser(*M*=3.86,*SD*=1.46), Misc.: autocompletion in MS Visual Studio, Eclipse IDE

At the end, participants were asked to rank the autocompletion interfaces based on their preference and to give reasons for their choices.

Task — Participants were given 24 tasks (3 tasks per interface). To simulate a realistic search task participants were asked to search and specify the birth place of a famous person (see Fig. 3). They were allowed to find the answers in Wikipedia and then fill in their answer using the autocompletion interface. Participants were encouraged to use autocompletion but could choose not to use it if they could not find the right suggestion from the list. We have chosen the non trivial tasks such as locations with exactly the same name. Thus, for all questions, the need for for disambiguation and choosing the correct terms was clear. For example, the birth place of Kurt Kobain (Aberdeen, Washington, USA) has at least 56 other similar place name matches, of which Aberdeen in the UK will most likely be the most familiar to our European participants. The time recorded are the autocompletion typing time only. We disregard the time it takes for the participant to browse the Web and look for answers.

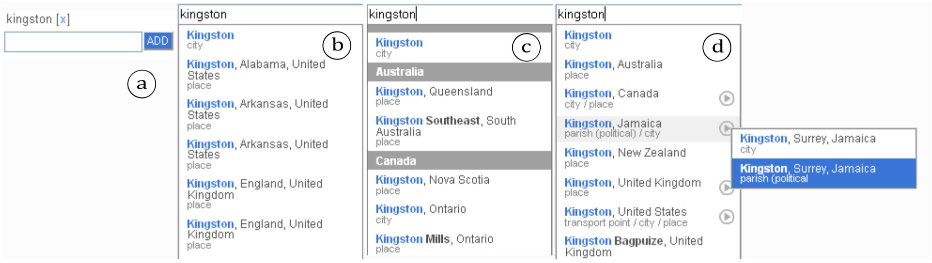
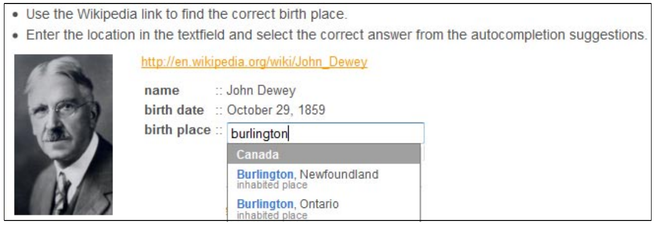

kingston [x]
ADD
a
kingston
b
Kingston city
Kingston, Alabama, United States place
Kingston, Arkansas, United States place
Kingston, Arkansas, United States place
Kingston, England, United Kingdom place
Kingston, England, United Kingdom place
kingston
c
Kingston city
Australia
Kingston, Queensland place
Kingston Southeast, South Australia place
Canada
Kingston, Nova Scotia place
Kingston, Ontario city
Kingston Mills, Ontario place
kingston
d
Kingston city
Kingston, Australia place
Kingston, Canada city / place
Kingston, Jamaica parish (political) / city
Kingston, New Zealand place
Kingston, United Kingdom place
Kingston, United States transport point / city / place
Kingston Bagpuize, United Kingdom
Kingston, Surrey, Jamaica city
Kingston, Surrey, Jamaica parish (political

Fig. 2. Autocompletion in Study 2, from left to right: a) NAC, b) Alphabetical order, c) Group by country and d) Composite.



- Use the Wikipedia link to find the correct birth place.
- Enter the location in the textfield and select the correct answer from the autocompletion suggestions.

http://en.wikipedia.org/wiki/John_Dewey

name :: John Dewey
birth date :: October 29, 1859
birth place :: burlington

Canada
Burlington, Newfoundland inhabited place
Burlington, Ontario inhabited place

Fig. 3. Task example used in Study 2

Results

• *Mean keystrokes:* As expected, an autocompletion interface reduces the number of keystrokes required to type. On average, users typed almost twice as many characters in the NAC condition (see Table 3). Additionally, we found that some participants copied and pasted the location name they found from Wikipedia. This behavior was identified from the interaction event log and is estimated to be about 7.5% from the total tasks performed by all participants.

• *Performance in Time:* Table 3 shows the mean time it took for participants to com-

Table 2. Quality of keywords provided by participants (492 tasks, 41 people, Study 2)

Interface	NAC	Alphabetical	Group	Composite
Total correct keyword	96.7%	86.2%	95.1%	84.5%
a. Unique concept	n/a	77.2%	86.2%	82.9%
b. One term	14.6%	2.4%	0.8%	0%
c. Two terms	53.7%	6.5%	5.7%	0%
d. Three terms	28.4%	0%	2.4%	1.6%
Total incorrect keyword	3.2%	13.8%	4.9%	15.4%
a. Choose wrong item	n/a	13.0%	4.9%	15.4%
b. Typing error	2.4%	0%	0%	0%
c. No keyword	0.8%	0.8%	0%	0%

Table 3. Comparison between NAC, Alphabetical, Group and Composite (492 tasks, n=41 people, Study 2)

Interface	NAC	Alphabetical	Group	Composite
Mean no of keystrokes (*SD*)	19.20(6.86)	8.55(4.50)	7.89(4.81)	7.91(3.82)
Mean time in ms (*SD*)	5943.61(3414.16)	38929.32 (46874.49)	18363.80 (10998.24)	17617.49 (12250.34)
Mean Rank (*SD*)	2.93(1.23)	2.71(.90)	1.98(1.11)	2.39(1.02)
Mean score Q1 * (*SD*)	3.07(1.21)	2.59(.87)	3.34(1.39)	3.56(.90)
Mean score Q2 * (*SD*)	n/a	3.05(1.24)	3.73(1.10)	3.61(.95)

* 5-Likert scale, score 1:strongly disagree, 5:strongly agree

plete a task. This time constituted the time from the first keystroke typed until selecting a suggestion (for the autocompletion conditions) or hitting the return key (for the NAC condition). When we compare the performance of the individual autocompletion interfaces, we find that Group and Composite are significantly faster (47% and 45% resp.) than the Alphabetical order[14] We conclude that both Group and Composite strategies help the user search for terms faster than Alphabetical order.

• *Quality of keywords:* Table 2 shows the quality of keywords provided by participants. The quality of keywords are measured by how accurate the location names are given. We found four levels of accuracy (from low to high): strings that consist of one term (mostly city names which are ambiguous because there exist many cities the same name e.g. *"Kingston"*), strings that consist of two terms (mostly a combination of city and state, or city and country e.g. *"Kingston, USA"*), strings that consist of three terms (mostly a combination of city, state and country names e.g. *"Kingston, Texas, USA"*) and keywords which were correctly chosen from the suggestions (unique concepts from the thesaurus). All autocompletion interfaces have a high percentage of correct keywords (all above 84.5%). The quality of keywords provided by participants, however, differ when using autocompletion and without. In NAC, most keywords consists of merely 2 terms (53.7%), which is in many cases not sufficient for location name disambiguation. For example, there are 47 places named *Kingston* in the USA. Only about a third of the cases in NAC (28.4%) consist of three terms. In contrast, keywords provided in the other autocompletion interfaces are mostly high quality keywords that are

[14] WSR *post-hoc* test result for Time: Group is sig. faster than Alphabetical ($p \ll .05$). Additionally, Composite is sig. faster than Alphabetical ($p \ll .05$).

unique concepts (86.2% for Alphabetical, 95.1% for Group, and 84.5% for Composite). We also identified three types of incorrect keywords: incorrect terms selected from the autocompletion suggestions, typing errors and blank entries where participants gave no keyword at all. A closer look at the incorrect keywords provided by participants reveals that most errors in the NAC conditions are typing mistakes (2.4%), while most errors in the autocompletion interfaces are wrong autocompletion selection. For example, selecting *Ottawa (the river)* instead of *Ottawa (the city)*. From all 3 autocompletion interfaces, Group organization generates least error (4.9%) compared to Composite (15.4%) and Alphabetical (13.8%). The results show that even though NAC is slightly faster, the quality of keywords provided in any of the autocompletion interfaces are far higher.

• *Perceived ease-of-use:* We gathered participants assessments on the ease-of-use of each interface (see Table 3). In general, people find the Group and Composite interface easier to use than the Alphabetical and NAC interface (χ^2(3)=17.52, $p \ll .05$)[15] In a follow up question (Q2), we wanted to know specifically if people understood the organization strategy. Most people agree with the statement that Group and Composite suggestion organization is easier to understand than Alphabetical list (χ^2(2)=8.12, p=.02)[16]. We conclude that both Group and Composite interfaces are perceived easier to use and understand than the Alphabetical order.

• *Preference:* Our analysis shows there is a preference for Group strategy (see Table 3), although Composite is not far behind (χ^2(3)=12.6, $p \ll .05$)[17]. From the comments made by the participants we understand more about the reasons behind the users preference. Participants acknowledge that autocompletion suggestions help avoid typing mistakes and enable them to express more keywords than they would otherwise have thought of. *"The lack of autocompletion choices prevents me to give a proper answer for question X."*[P1]. In general, participants think Group organization is better. *"It's comfortable to see the countries separated"* [P2]. *" You know where you have to go. You get a better overview"* [P16]. For many participants, the Composite organization is relatively new. The main disadvantages of Composite are: a) requires more interaction with the interface before making a selection (e.g. mouse movement and click) and b) submenu interaction requires getting used to *"took several seconds to discover the small arrows. After that, the interface is easy to use"* [P4]

4 Discussion

Study limitation — In Study 1, the grouping strategies tested were developed partly by trial and error in combination with educated guesses. We came up with a range of possible grouping strategies, which we tested informally. We only formally tested the three different grouping strategies which performed best in the informal test. The result of Study 1 shows that, in contrary to the TGN grouping strategies, our best grouping

[15] WSR *post-hoc* test result for Q1: Group is sig. perceived easier-to-use than Alphabetical ($p \ll .05$). Composite is sig. perceived easier-to-use then Alphabetical ($p \ll .05$). No difference between Group and Composite.

[16] WSR *post-hoc* test result for Q2: Group organization is sig. perceived easier to understand than Alphabetical (p=.01). Composite organization is sig. perceived easier to understand than Alphabetical (p=.04). No difference between Group and Composite.

[17] WSR *post-hoc* test result for preferred interface: Group organization is sig. preferred than Alphabetical ($p \ll .05$) and NAC ($p \ll .05$). No sig. difference between Group and Composite.

strategies for WordNet were not helpful for users and people prefered the Alphabetical order as an organization strategy. We acknowledge that it might be the case that we did not succeed to find the appropriate grouping for WordNet. Therefore it is reasonable to only conclude that for the WordNet grouping strategies which were tested, none outperformed the Alphabetical list.

During the experiment, we observed that there are some cut and paste behavior specially for the NAC interface. This might provide additional explanation as to why participants complete NAC tasks faster than other tasks. We expect that if all participants were only allowed to type (not cut and paste), the autocompletion interfaces would show comparable time performance.

Alphabetical order — When using a global thesaurus, such as WordNet, Alphabetical order organization seems to be the best option. This organization requires very little learning effort. The downside of this organization is, as one participant points out, that it provides no “overview” when there are many suggestions.

Grouping strategy — We learned from the first study that a grouping strategy should be chosen carefully because not every grouping strategy is suitable to use. The TGN groupings produced by the thesaurus hierarchy seem to be more natural than WordNet groupings. In many of our pairwise statistical comparison between Group and Composite organization, we found no significant differences. Based on this study alone, we cannot see a clear advantage one type of organization over the other. We can say, however, that the Group organization has a tendency for breadth(expanding the length of suggestion interface vertically), whereas the Composite organization has a tendency to shorten the breadth of the suggestion interface. Therefore, depending on the thesaurus used and the length of suggestions it produces, the Composite organization might have an advantage.

Improve autocompletion — The server log indicates that some people use commas and make keystroke errors. We learned that in order to make a good autocompletion interface, there are a number of supporting functionalities that are indispensable: (a) Compensate for non alphanumeric letters such as white space(s) and commas. For example, the system should know that *Kingston Jamaica* is the same query as *Kingston, Jamaica*. Our finding is consistent with the study in [15] on how people express similar queries in different ways. (b) Users may make typing mistakes (e.g. *Ottawa, Ottowa, Otawa*). Spell check and giving suggestions based on likely spelling would be a useful feature.

Incentive to use autocompletion — We observe that autocompletion can stimulate people to be more specific in their keywords. Even though participants were instructed to be ”as precise as possible”, the keywords provided in the no-autocompletion tasks are largely ambiguous: only less than a third of the keywords consist of detail information (i.e. city, state and country). This is in large contrast to the keywords provided by the participants when using autocompletion suggestions where they are mostly unambiguous location concepts. Autocompletion allows people to provide high quality keywords from which an information retrieval system can benefit. We believe that users are willing to spend more time to formulate a more elaborate query with the help of autocompletion interfaces if the option to use is made available.

5 Conclusions and Future work

We conducted serial user studies to compare different kinds of organization for autocompletion suggestions that can help improve known-item search task. In the first study, we found that grouping strategies might not be suitable when using a global thesauri, such as WordNet and only certain grouping strategies could be used for TGN. Based on what we have learned, we conducted a second study where we compared three different autocompletion suggestion interfaces. In general, we found that the quality of keywords provided by users are better with the autocompletion; Group and Composite organization help users search faster than when using the Alphabetical order; users perceive Group and Composite easier to use and to understand. We are currently integrating autocompletion with our applications and evaluate its performance for a domain-expert annotation task. In addition to this, we will improve autocompletion interface to detect similar query strings identified in [15], such as synonyms, extra whitespace and word swaps.

6 Acknowledgments

This research was supported by the MultimediaN project through the BSIK program of the Dutch Government and by the European Commission contract FP6-027026, K-Space.

References

1. E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM.
2. anonymous for review.
3. H. Bast, D. Majumdar, and I. Weber. Efficient interactive query expansion with complete search. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 857–860, New York, NY, USA, 2007. ACM.
4. M. Beaulieu. Experiments on interfaces to support query expansion. *Journal of Documentation*, 53:8—19, 1997.
5. E. N. Efthimiadis. Interactive query expansion: a user-based evaluation in a relevance feedback environment. volume 51, pages 989–1003, New York, NY, USA, 2000. John Wiley & Sons, Inc.
6. J. I. B. Gonzales. A theory of organization. In *SIGDOC '94: Proceedings of the 12th annual international conference on Systems documentation*, pages 145–155, New York, NY, USA, 1994. ACM.
7. J. J. Hendrickson. Performance, preference, and visual scan patterns on a menu-based system: implications for interface design. In *CHI '89: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 217–222, New York, NY, USA, 1989. ACM.
8. G. M. Hodgson and S. R. Ruth. The use of menus in the design of on-line sytems: a retrospective view. *SIGCHI Bull.*, 17(1):16–22, 1985.

9. M. Jakobsson. Autocompletion in full text transaction entry: a method for humanized input. In *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 327–332, New York, NY, USA, 1986. ACM.
10. H. Joho, C. Coverson, M. Sanderson, and M. Beaulieu. Hierarchical presentation of expansion terms. In *SAC '02: Proceedings of the 2002 ACM symposium on Applied computing*, pages 645–649, New York, NY, USA, 2002. ACM.
11. H. Joho, M. Sanderson, and M. Beaulieu. A study of user interaction with a concept-based interactive query expansion support tool. In *ECIR 2004*, pages 42–56, 2004.
12. E. S. Lee and D. R. Raymond. Menu-driven systems. In A. Kent and J. G. Williams, editors, *Encyclopedia of Microcomputers*, volume 11, pages 101–127. Marcel Dekker, New York, 1993.
13. F. Radlinski. Query chains: Learning to rank from implicit feedback. In *In KDD*, pages 239–248. ACM Press, 2005.
14. E. Rosch. Principles of categorization. *Cognitive Science, a Persepective from Psychology and Artificial Intelligence*, 1988.
15. J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158, New York, NY, USA, 2007. ACM.
16. R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166, New York, NY, USA, 2007. ACM.
17. R. W. White and G. Marchionini. A study of real-time query expansion effectiveness. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 715–716, New York, NY, USA, 2006. ACM.
18. R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43(3):685–704, 2007.
19. R. W. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *SIGIR '07: Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 255–262, New York, NY, USA, 2007. ACM Press.