Organizing Suggestions in Autocompletion Interfaces[⋆]

Alia Amin[1], Michiel Hildebrand[1], Jacco van Ossenbruggen[1,4],
Vanessa Evers[2], and Lynda Hardman[1,2,3]

[1] CWI, NL
[2] University of Amsterdam, NL
[3] Eindhoven University of Technology, NL
[4] VU University, NL

Abstract. We describe two user studies that investigate organization strategies of autocompletion in a known-item search task: searching for terms taken from a thesaurus. In Study 1, we explored ways of grouping term suggestions from two different thesauri (TGN and WordNet) and found that different thesauri may require different organization strategies. Users found *Group* organization more appropriate for location names from TGN, while *Alphabetical* works better for object names from WordNet. In Study 2, we compared three different organization strategies (*Alphabetical*, *Group* and *Composite*) for location name search tasks. The results indicate that for TGN autocompletion interfaces help improve the quality of keywords, *Group* and *Composite* organization help users search faster, and is perceived easier to understand and to use than *Alphabetical*.

1 Introduction

Interactive query expansion (IQE) is being researched as a means to help improve user search performance and query quality. Real-time query expansion (RTQE), such as autocompletion, is adopted in many search applications e.g. *Google Suggest* or *Yahoo! Search Assist*. Most research efforts are directed towards improving query expansion suggestions, e.g.[1,4,8,10], and generally pay less attention to the interface issues. Many RTQE interfaces use only *list* organization as a presentation style. Prior work has, however, led us to believe that different types of implementation of RTQE presentation would likely result in different user search performance. In [3], three different interfaces to the same retrieval system were compared. The study suggests that the quality and effectiveness of search depends on how well the retrieval system and its interface support query expansion. Joho et al. [6] compared two types of query expansion presentation styles: alphabetical *order* and *menu hierarchy* and found that even though there is no significant difference in the precision-recall between using the two interfaces, people finished the search task significantly faster when using the *menu hierarchy*. Another study [7] compared two variants of hierarchical IQE system against a baseline and found that the hierarchies reduce search iterations and paging actions, and increase the chance to find relevant items.

In this research, we focus on the presentation aspects of autocompletion, namely organization strategies and how they influence users' search performance. We are motivated

[⋆] This research was supported by the MultimediaN project through the BSIK program and by the European Commission contract FP6-027026, K-Space.

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 521–529, 2009.

by the usage of relationships of terms from a thesaurus to improve RTQE presentation. Certain relationships between terms from a thesaurus have been known to improve the quality of query expansion. Efthimiadis et al. [4] investigated the terms used in an 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, in [6], the most useful relationships for WordNet are hyponym, hypernym and synonym. In this study, 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 faster and easier. We carried 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 influence the search quality and performance.

2 Organization of Suggestions

Fig.1 shows different organization strategies for autocompletion suggestions were taken from TGN[1] autocompletions. Similar visualizations and algorithms were applied to WordNet.

Alphabetical order — Fig.1a shows autocompletion suggestions in alphabetical ordering. The location name "*Kingston*, Alabama" is shown before "*Kingston*, Arkansas". Exact matches are presented first, followed by partial matches.

Group — In Fig 1b and c, a group category is conveyed visually under one group title. Where terms are related by explicit thesaurus relations, any of these relations can be used as a basis for grouping e.g. 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 > United Kingdom > Kingston*). Grouping can be based on any predefined level within this hierarchy, e.g. grouping by country (Fig. 1b), or based on a common property, such as place type (Fig. 1c) e.g. inhabited place (city, village) or body of water (stream, lake).

In the dynamic grouping, the group headings are determined by an algorithm that optimizes groups based on the number of suggestions retrieved and their relative positions in the thesaurus structure. In Dynamic TD, the grouping algorithm traverses the thesaurus structure top down to group the suggestions. In Dynamic BU, this is done bottom up. Dynamic groups could provide an alternative grouping for thesauri with irregular hierarchical structures such as WordNet.

Composite — A composite organization resembles a two level cascaded menu hierarchy. In Fig. 1d, the primary menu contains all exact matches of all location names from the same country. The submenu displays more information about the location names that allows disambiguation e.g. *Kingston (the city)* or *Kingston (the parish)*. This strategy retains the simplicity of alphabetical order, while giving access to larger numbers of alternatives in the same screen real-estate.

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

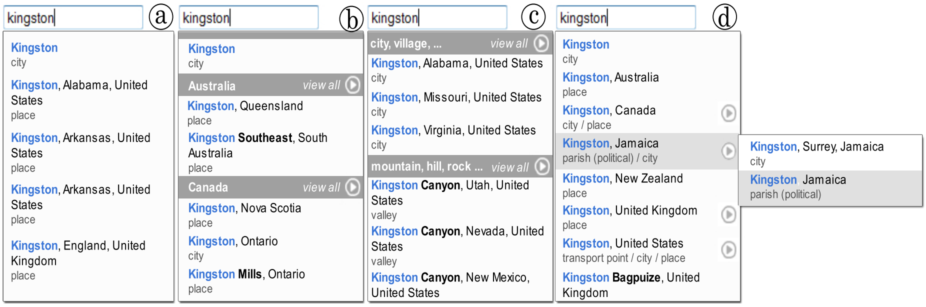


Fig. 1. Different autocompletion organization strategies used in Study 1 and 2 for TGN, a) Alphabetical, b) Group by Country, c) Group by Place type, d) Composite

3 User Studies

We conducted two user studies: the first to explore the effects of grouping strategies on two different thesauri. The second to investigate different organization strategies on the same set of suggestions. See [2] for a detailed description of both studies and [5] for the autocompletion architecture.

3.1 Study 1: Grouping Strategies

The goal of Study 1 is to investigate to what extent grouping strategies for autocompletion suggestions can be applied to thesauri and if appropriate, which grouping strategies are meaningful for users. We chose to implement similar grouping strategies for two different thesauri: a geographical thesaurus, TGN and a lexical thesaurus, WordNet. Our intention was not to compare the two thesauri, but to evaluate the suitability of different grouping strategies when implemented for these thesauri.

Interfaces — In total, 8 autocompletion interfaces were selected for TGN and WordNet after informal trials and selection from different algorithms and combinations. The groupings for TGN location names (LN) are: by country (Fig. 1b), by place type (Fig. 1c) and Dynamic TD. Alphabetical (Fig. 1a) is used as a baseline. For WordNet object names (ON), the 3 grouping strategies are: predefined grouping using the top nine WordNet category nouns from the hypernym hierarchy, and two dynamic groupings: Dynamic TD and Dynamic BU. Alphabetical is, also, used as a baseline.

Participants — Participants were recruited from universities and institutes from diverse departments, such as computer science and natural science. Participants (47 people) were mostly students and some university employees. All participants use the Internet daily and are familiar with the autocompletion (e.g. in email clients and search engines), 14 participants used advanced autocompletion e.g. in script or source code editors.

Procedure and Tasks — The study was an online interactive experiment. All session activities were logged. First, participants answered a short questionnaire about their autocompletion experiences. Next, every participant was assign tasks with 4 TGN-LN interfaces (within subject design). For each LN interface, participants were given the same tasks: to formulate several location queries, such as *Berlin* (city name), and find the correct location names from the autocompletion suggestions. After completing the tasks, participants were asked to answer assessment questions about the quality of the groupings and to give comments. Finally, participants were asked to rank their preferred LN interface, from the most to the least preferred, and provide reasons. The same procedure is repeated by the participants for the WordNet-ON interfaces. The assessment criteria on the quality of grouping are derived from references in [2]. These criteria questions are:

Q1 - *"I think the items belonging to each group in this list are similar to each other."*

Q2 - *"I think the items belonging to different groups in this list are different from each other."*

Q3 - *"I think the relationship between the items and group title is clear in this list."*

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

Q5 - *"I think the titles of the groups in this list are clear."*

Results

• *Assessment:* Table 1(left) shows participants' mean assessment scores for LN and ON grouping strategies. We use Friedman two-way analysis by ranks (FTWAR)[2] to analyze each assessment criteria. For LN, we found that: (a) Place type grouping scored best with respect to perceived similarity - Q1 (χ^2(2)=7.36, *p*=.03)[3]. Perceived similarity indicates the cohesiveness between the suggestions in a group. (b) Country grouping scored best with respect to group title appropriateness - Q5 (χ^2(2)=6.77, *p*=.03)[4] (c) Country grouping scored lowest with respect to the number of groups - Q4 (χ^2(2)=8.11, *p*=.02) [5] The Country group strategy gives most representative group titles (Q5) but scores poor on the number of group (Q4).

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.

For the ON interfaces: (a) Dynamic BU group scored lowest with respect to perceived difference - Q2 (χ^2(2)=10.17,*p*=.01)[6] (b) Dynamic TD group scored lowest with respect to the number of groups - Q4 (χ^2(2)=9.66, *p*=.01)[7]. The results showed that none of the ON group strategies excels from each other in the assessment score. We only found that the Dynamic TD and Dynamic BU groups perform the worst in Q2 and Q4. We think this is because the dynamic group strategies actually add to participants' cognitive burden when they are trying to go through the suggestion list and understand the different categories every time. No grouping strategy in ON is assessed the best by our participants.

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

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

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

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

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

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

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

TGN-LN	Mean Score (*SD*) *			
Question	Place type	Country	Dynamic TD	*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 TD	Dynamic BU	*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 TD	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 TD	2.62(.97)	
Predefined	2.68(1.09)	
Dynamic BU	2.72(1.06)	

• *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 FTWAR, we found no strong preference in any LN interfaces. (χ^2(3)=5.14, p>.05). The comments provided by the participants indicate that they prefer different interfaces for different reasons. We conducted the same analysis for the four ON interfaces and found a different result. Participants strongly preferred Alphabetical to all other organization strategies ($\chi^2(3)=10.38, p=.02$)[8]. Many participants commented that it is difficult to understand the ON grouping strategies, which led to a strong preference for Alphabetical.

Retrospective — In Study 1, we wanted to find out 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 is the one that scores highest on all five assessment scores (Table 1 left) and most preferred (Table 1 right). This is, however, not the case. For TGN, different groupings are favored in different ways. We could find a sensible grouping strategy, e.g. by country or by place type, that people could understand relatively easily. For WordNet, however, the results of the users preference and assessment scores led to the conclusion that the group organization should not be used. In cases where the underlying thesaurus does not provide the information necessary for appropriate grouping, the Alphabetical is the best option.

3.2 Study 2: Organization Strategies

The goal of Study 2 is to compare 3 types of autocompletion for TGN: Alphabetical, Group and Composite. We decided not to use WordNet because none of the group strategies offered for WordNet in Study 1 outperformed the baseline (Alphabetical). Users are required to use autocompletion for known-item search tasks. We measure search

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

performance (time to complete task and quality of keywords) and ease-of-use (users' assessments and preference).

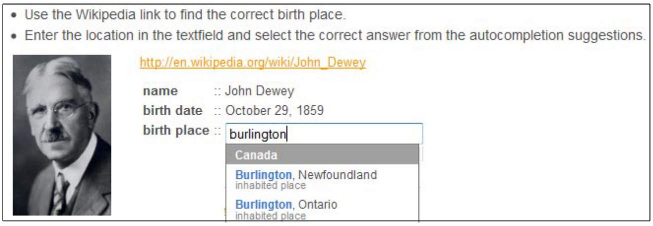
Interfaces — We compared 4 different interfaces, namely: Alphabetical (Fig.1a), Group (Fig.1b), Composite (Fig.1d) and no autocompletion (NAC) interface.

Participants — We recruited 41 participants in the same manner as for Study 1.

Procedure — Each participant was assigned 4 interfaces: NAC, Alphabetical, Group and Composite (within subject design). Participants started by answering general questions about their autocompletion experience. Afterwards, participants were given 12 tasks. In every task, time measurements were taken. After every interface, participants answered two questions about the usability of the different interfaces(5-Likert scale):

Q1 - *"I find this interface easy to use."*
Q2 - *"I find the organization of the suggestions easy to understand.*"

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



- 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. 2. Task example used in Study 2

Task — Participants were given 12 tasks (3 tasks per interface). Participants were asked to search and specify the birth place of a famous person (see Fig. 2). They were allowed to find the answers in Wikipedia and fill them in using the autocompletion interface. Participants could choose not to use autocompletion and type the answers manually if they could not find the right suggestion.

Results

• *Performance in time:* The mean time it took for participants to complete a task, which is the time from the first keystroke typed until selecting a suggestion (for the autocompletion conditions) or hitting the return key (for the NAC condition). We disregard the time it took for participants to browse the Web and look for answers. In general, users used more than twice as many keystrokes in NAC compared to when using autocompletion (see Table 3). When comparing the performance of individual autocompletion interfaces, we find that Group and Composite are significantly faster (47% and 45% resp.) than Alphabetical[9]. We conclude that both Group and Composite strategies help the user search for terms faster than Alphabetical.

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

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. Select wrong item	n/a	13.0%	4.9%	15.4%
b. Typing error	**2.4%**	0%	0%	0%
c. No answer	0.8%	0.8%	0%	0%

Table 3. User search performance and preference (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 s (*SD*)	5.94(3.41)	38.93 (46.87)	18.36 (10.99)	17.62 (12.25)
Mean pref. 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

• *Quality of keywords:* Table 2 shows the quality of keywords provided by participants. The quality of keywords is measured by how accurately the location names are given. We identified 3 types of errors: incorrect terms selected from the autocompletion suggestions, typing errors and missing keyword (no answer). Most NAC errors came from typing mistakes (2.4%), while in the autocompletion interfaces, they came from wrong autocompletion selection, e.g. selecting *Ottawa (the river)* instead of *Ottawa (the city)*. For the correct keywords, we found 4 levels of accuracy (from low to high): one term strings (mostly city names, e.g. *"Kingston"*), two terms strings (mostly city and state/country, e.g. *"Kingston, USA"*), and three terms strings (mostly city, state and country names, e.g. *"Kingston, Texas, USA"*) and keywords which are unique concepts from the thesaurus. The quality of keywords provided differs with and without autocompletion. In NAC, most keywords consist of merely 2 terms (53.7%), which is in many cases insufficient for disambiguation, e.g. there are 47 places named *Kingston* in the *USA*. In contrast, keywords provided in the other autocompletion interfaces are mostly high quality keywords that are unique concepts (86.2% Alphabetical, 95.1% Group, and 84.5% Composite). The results show that the quality of keywords provided by Autocompletion interfaces are far better.

• *Perceived ease-of-use and preference:* In general, people find the Group and Composite interface easier to use than Alphabetical and NAC interface (for Q1 $\chi^2(3)=17.52$, $p \ll .05$)[10] (see Table 3). In a follow-up question (Q2), we found that most people

[10] 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.

think that Group and Composite suggestion organization is easier to understand than Alphabetical list ($\chi^2(2)$=8.12, p=.02)[11]. Moreover, Table 3 shows Group strategy and Composite is most preferred ($\chi^2(3)$=12.6, $p \ll .05$)[12]. We conclude that both Group and Composite interfaces are perceived easier to use and to understand than Alphabetical.

4 Discussion and Conclusion

Alphabetical order — For a domain independent lexical thesaurus, such as WordNet, Alphabetical order seems to be the best option. Alphabetical order requires very little learning effort. The downside of this organization is that it provides no "overview" when there are many suggestions.

Grouping strategy — Study 1 showed that a grouping strategy should be chosen carefully because not every grouping strategy is suitable. The TGN groupings based on the geographical hierarchy seem to make more sense than the WordNet groupings based on the domain independent lexical hierarchy. In many of our pairwise statistical comparisons between Group and Composite organization, we found no significant differences. The Group organization, however, tends to expand the length of the suggestions interface vertically, whereas the Composite organization tends to expand horizontally using submenus. Therefore, depending on the thesaurus used and the length of suggestions it produces, the Composite organization might have an advantage.

Autocompletion improvements — In order to make a well designed autocompletion interface, several supporting functionalities are indispensable:

(a) Compensate for non alphanumeric characters, such as white space(s) and commas. For example, the system should know that *Kingston - Jamaica* is the same query as *Kingston, Jamaica*. This finding is consistent with [9] on how people express similar queries in different ways. (b) Spell check to avoid typing mistakes and provide likely suggestions (e.g. *Ottawa, Ottowa, Otawa*). (c) Detect similar query strings identified in [9], such as synonyms and word swaps.

References

1. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proc. SIGIR 2006, pp. 19–26 (2006)
2. Amin, A., Hildebrand, M., van Ossenbruggen, J.R., Evers, V., Hardman, L.: Organizing Suggestion. In: Autocompletion Interfaces. INS-E0901, CWI (2009)
3. Beaulieu, M.: Experiments on interfaces to support query expansion. Journal of Documentation 53, 8–19 (1997)
4. Efthimiadis, E.N.: Interactive query expansion: a user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science 51, 989–1003 (2000)

[11] 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.

[12] 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.

5. Hildebrand, M., van Ossenbruggen, J.R., Amin, A.K., Aroyo, L., Wielemaker, J., Hardman, L.: The Design Space Of A Configurable Autocompletion Component. INS-E0708, CWI (2007)
6. Joho, H., Coverson, C., Sanderson, M., Beaulieu, M.: Hierarchical presentation of expansion terms. In: Nyberg, K., Heys, H.M. (eds.) SAC 2002. LNCS, vol. 2595, pp. 645–649. Springer, Heidelberg (2003)
7. Joho, H., Sanderson, M., Beaulieu, M.: A study of user interaction with a concept-based interactive query expansion support tool. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 42–56. Springer, Heidelberg (2004)
8. Radlinski, F.: Query chains: Learning to rank from implicit feedback. In: Proc. KDD, pp. 239–248 (2005)
9. Teevan, J., Adar, E., Jones, R., Potts, M.A.S.: Information re-retrieval: repeat queries in Yahoo's logs. In: Proc. SIGIR 2007, pp. 151–158 (2007)
10. White, R.W., Bilenko, M., Cucerzan, S.: Studying the use of popular destinations to enhance web search interaction. In: Proc. SIGIR 2007, pp. 159–166 (2007)