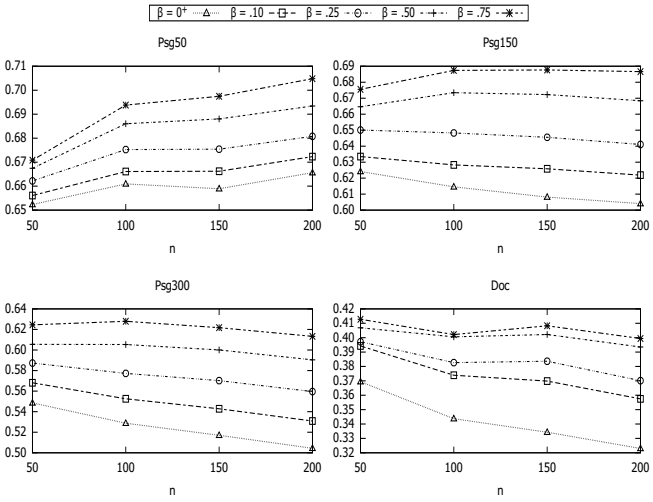

Figure 1: The effect of n on the Focused-Focused test for INEX with $k = 4$. Note: graphs are not to the same scale.

Table 4: Graded-Binary and Graded-Graded tests performed for documents. '1' marks statistically significant difference with respect to "$\beta = 1$".

		Graded-Binary				Graded-Graded			
		GOV2		ClueWeb		GOV2		ClueWeb	
n	k	$\beta = 1$	$\beta = 2$	$\beta = 1$	$\beta = 2$	$\beta = 1$	$\beta = 2$	$\beta = 1$	$\beta = 2$
50	4	.737	$.771^1$	.641	.656	1.001	1.123^1	.952	1.017^1
	9	.673	$.703^1$	.581	.598	.908	$.975^1$	.859	$.890^1$
100	4	.710	$.762^1$	.606	.631	.950	1.135^1	.881	$.997^1$
	9	.657	$.691^1$	.540	.561	.875	$.987^1$	.784	$.858^1$
150	4	.709	$.755^1$	.580	.597	.934	1.130^1	.829	$.938^1$
	9	.650	$.687^1$	.506	.525	.854	$.986^1$	.723	$.800^1$
200	4	.699	$.752^1$	.558	.585	.911	1.121^1	.798	$.930^1$
	9	.644	$.683^1$	.487	.512	.837	$.978^1$	.697	$.789^1$

document lists for the TREC datasets. These datasets have no focused relevance judgments; therefore, the tests are performed only for documents. Furthermore, the INEX dataset is not used as it has no graded relevance judgments. The queries used are those whose retrieved list contains at least one relevant seed with $\beta = 2$ and an additional relevant document. Since the comparisons we discuss are for the same list (determined by n), the test results are comparable and there is no need to use the ratio values as in Table 2.

Table 4 shows that in all cases, increasing the relevance degree (grade) of the seed document results in its neighborhood containing more relevant documents (Graded-Binary) with higher relevance grades (Graded-Graded). These findings are conceptually reminiscent of those presented above for the Focused-Binary and Focused-Focused tests. We also see that the increase in the tests' values when moving from $\beta = 1$ to $\beta = 2$ is always statistically significant for GOV2; for ClueWeb, statistically significant increase is observed for the Graded-Graded test.

5 CONCLUSIONS AND FUTURE WORK

We presented novel cluster hypothesis tests that utilize graded and focused relevance judgments. We found that (i) the cluster hypothesis holds for passages; (ii) relevant items (documents or passages) that contain a high fraction of relevant text are more similar to other relevant items (specifically, those with a high fraction of relevant text) than relevant items with low fraction; and, (iii) documents marked as highly relevant are more similar to other relevant documents (specifically, highly relevant) than relevant documents not marked as such. These findings motivate the development of passage retrieval methods that utilize inter-passage similarities.

Acknowledgments. We thank the reviewers for their comments. This paper is based upon work supported in part by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

REFERENCES

[1] Paavo Arvola, Shlomo Geva, Jaap Kamps, Ralf Schenkel, Andrew Trotman, and Johanna Vainio. 2011. Overview of the INEX 2010 ad hoc track. In *Comparative Evaluation of Focused Retrieval*. Springer, 1–32.
[2] James P. Callan. 1994. Passage-Level Evidence in Document Retrieval. In *Proc. of SIGIR*. 302–301.
[3] David Carmel, Anna Shtok, and Oren Kurland. 2013. Position-based contextualization for passage retrieval. In *Proc. of CIKM*. 1241–1244.
[4] Gordon V Cormack, Mark D Smucker, and Charles LA Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval* 14, 5 (2011), 441–465.
[5] Abdelmoula El-Hamdouchi and Peter Willett. 1987. Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science* 13, 6 (1987), 361–365.
[6] Ronald T. Fernández, David E. Losada, and Leif Azzopardi. 2011. Extending the language modeling framework for sentence retrieval to include local context. *Information Retrieval* 14, 4 (2011), 355–389.
[7] Norbert Fuhr, Marc Lechtenfeld, Benno Stein, and Tim Gollub. 2012. The optimum clustering framework: implementing the cluster hypothesis. *Information Retrieval* 15, 2 (2012), 93–115.
[8] Shlomo Geva, Jaap Kamps, Miro Lethonen, Ralf Schenkel, James A Thom, and Andrew Trotman. 2010. Overview of the INEX 2009 ad hoc track. In *Focused retrieval and evaluation*. Springer, 4–25.
[9] Nick Jardine and C. J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7, 5 (1971), 217–240.
[10] Mostafa Keikha, Jae Hyun Park, W Bruce Croft, and Mark Sanderson. 2014. Retrieving passages and finding answers. In *Proc. of ADCS*. 81.
[11] Oren Kurland. 2009. Re-ranking search results using language models of query-specific clusters. *Information Retrieval* 12, 4 (2009), 437–460.
[12] Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, and Frédéric Saubion. 2008. Using text segmentation to enhance the cluster hypothesis. In *Proc. of AIMSA*. 69–82.
[13] Vanessa Murdock and W Bruce Croft. 2005. A translation model for sentence retrieval. In *Proc. of HLT-EMNLP*. 684–691.
[14] S.-H. Na, I.-S. Kang, and J.-H. Lee. 2008. Revisit of nearest neighbor test for direct evaluation of inter-document similarities. In *Proc. of ECIR*. 674–678.
[15] Fiana Raiber and Oren Kurland. 2012. Exploring the cluster hypothesis, and cluster-based retrieval, over the Web. In *Proc. of CIKM*. 2507–2510.
[16] Fiana Raiber and Oren Kurland. 2013. Ranking document clusters using markov random fields. In *Proc. of SIGIR*. 333–342.
[17] Fiana Raiber and Oren Kurland. 2014. The correlation between cluster hypothesis tests and the effectiveness of cluster-based retrieval. In *Proc. of SIGIR*. 1155–1158.
[18] Fiana Raiber, Oren Kurland, Filip Radlinski, and Milad Shokouhi. 2015. Learning asymmetric co-relevance. In *Proc. of ICTIR*. 281–290.
[19] Hadas Raviv, Oren Kurland, and David Carmel. 2013. The cluster hypothesis for entity oriented search. In *Proc. of SIGIR*. 841–844.
[20] Mark D Smucker and James Allan. 2009. A new measure of the cluster hypothesis. In *Proc. of ICTIR*. 281–288.
[21] Anastasios Tombros, Robert Villa, and C. J. Van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information processing & management* 38, 4 (2002), 559–582.
[22] C. J. van Rijsbergen. 1979. *Information Retrieval* (second ed.). Butterworths.
[23] Ellen M. Voorhees. 1985. The cluster hypothesis revisited. In *Proc. of SIGIR*. 188–196.
[24] Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of SIGIR*. 334–342.