dr Marcin Sydow
IPI PAN oraz Web Mining Lab PJWSTK, Warszawa
E-mail: msyd@ipipan.waw.pl

Różnorodność wyników w wyszukiwaniu informacji

W problemie wyszukiwania informacji (ang. *information retrieval*, [MRS08]) mamy do czynienia ze zbiorem D dokumentów (tzw. korpus) i zapytaniem q (ze zbioru możliwych zapytań Q), które wyraża tzw. *potrzebę informacyjną* u użytkownika. Zbiór wszystkich możliwych zapytań Q jest zdefiniowany za pośrednictwem składni zapytań. Zadaniem systemu jest zwrócić zbiór S składający się z k (zwykle $k \ll |D|$) najbardziej „*trafnych*" (ang. *relevant*) dokumentów w kontekście zapytania q. Jednym z głównych problemów praktycznych jest to, że faktyczna potrzeba informacyjna użytkownika jest systemowi wyszukującemu nieznana, a jest jedynie w sposób niedoskonały reprezentowana przez zapytanie q, które może mieć bardzo *różne interpretacje* (np. zapytanie „zamki"). W dominującym dziś modelu (używanym np. w wyszukiwarkach WWW), w pewnym uproszczeniu, *trafność* jest modelowana za pomocą nieujemnej funkcji rzeczywistej $\mathrm{REL} : D \times Q \to R^+$, gdzie im wyższa wartość $\mathrm{REL}(d, q)$, tym wyższa trafność dokumentu. W modelu tym, wywodzącym się z tzw. założenia PRP (ang. *Probability Ranking Principle* [Rob77]) przy danym zapytaniu q wartość funkcji trafności $\mathrm{REL}(d, q)$ obliczana jest dla każdego dokumentu d osobno i system zwraca k dokumentów o największej wartości trafności (posortowanych nierosnąco po tej wartości). Model ten zakłada, że trafność dokumentu d jest niezależna od trafności innych dokumentów (ang. *independent relevance assumption*), co upraszcza obliczenia, ale powoduje też pewne problemy:

1) k zwróconych dokumentów może reprezentować wysoce powtarzalną informację,
2) wyniki w takim modelu mogą zostać zdominowane przez interpretację zapytania odpowiadającą najliczniejszej grupie użytkowników, co sprawia, że pozostali użytkownicy nie znajdą *żadnego* trafnego dokumentu wśród zwróconych wyników.

Rozwiązaniem tego problemu jest stosunkowo nowe podejście polegające na *dywersyfikacji* wyników wyszukiwania, gdzie przy obliczaniu trafności dokumentu uwzględnia się też pozostałe zwracane dokumenty, tak aby zwrócone wyniki łącznie reprezentowały jak najwięcej możliwych interpretacji czy aspektów danego zapytania.

Przedstawiona zostanie przykładowa miara trafności uwzględniająca dywersyfikację [AGHI09], której optymalizacja stanowi, jak pokazano, problem NP-trudny (redukcja z problemu MAX-COVERAGE), która dzięki własności submodularności umożliwia jednak zastosowanie prostego zachłannego algorytmu aproksymacyjnego o stałym współczynniku aproksymacji $(1 - 1/e)$ [Hoc96]. Następnie przedstawiony zostanie nowy model reprezentacji wiedzy i zapytań za pomocą tzw. *semantycz-*

nych grafów wiedzy, który umożliwia tzw. *wyszukiwanie semantyczne*, będące poza zasięgiem możliwości dzisiejszych wyszukiwarek WWW, a stanowi przyczynek do nowej generacji wyszukiwarek. Semantyczne grafy wiedzy są multigrafami, gdzie wierzchołki reprezentują encje, a skierowane krawędzie reprezentują binarne relacje między nimi. W szczególności przedstawiony zostanie zaproponowany niedawno przez autora problem DIVERSUM [SPS10], w którym dla danego węzła q (encji) grafu należy zwrócić spójny podgraf S semantycznego grafu wiedzy zawierający q, co najwyżej k krawędzi i stanowiący „zdywersyfikowane" podsumowanie informacji o encji q [SPS+10]. Problem ten w naturalny sposób wiąże się z zagadnieniem dywersyfikacji opisanym wcześniej w zupełnie nowym medium, którym są semantyczne grafy wiedzy. Sformułowany zostanie główny temat planowanych badań autora polegający na zdefiniowaniu odpowiedniej „funkcji trafności" dla semantycznych grafów wiedzy uwzględniający dywersyfikację wyników, analizie złożoności optymalizacji tej miary i poszukiwaniu efektywnych algorytmów rozwiązujących ten problem.

Literatura

[AGHI09] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, *Diversifying search results*, in: WSDM '09, Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, New York 2009, pp. 5–14.

[Hoc96] D. Hochbaum, ed., *Approximation Algorithms for NP-hard Problems*, PWS Publishing Company, 1996.

[MRS08] C. D. Manning, P. Raghavan, H. Schutze, *An Introduction to Information Retrieval*, Cambridge Univ. Press, 2008.

[SPS10] M. Sydow, M. Pikuła, R. Schenkel, *DIVERSUM: towards diversified summarisation of entities in knowledge graphs*, Data Engineering Workshops (ICDEW), 2010 IEEE 26th ICDE Conference, IEEE, 2010, pp. 221–226.

[SPS+10] M. Sydow, M. Pikuła, R. Schenkel, A. Siemion, *Entity summarisation with limited edge budget on knowledge graphs* (accepted for Computational Linguistics - Applications, CLA 2010 Workshop), 2010.

[Rob77] S. Robertson, *The probability ranking principle*, Journal of Documentation 33:4 (1977), pp. 294–304.