- Java – это кофе.

Очевидно, что без использования истории запроса невозможно догадаться о значении термина “Java”, поэтому история запросов является важным компонентом.

Допустим, в истории часто встречается программирование, поэтому к запросу можно привязать понятие «Java – это язык программирования». Пусть в некотором абзаце встречается термин “Java”, если в этом абзаце также встречаются компьютерные термины, то к абзацу на этапе индексирования будет привязано понятие «Java – это язык программирования». В этом случае мы найдем по запросу все абзацы, связанные с языком программирования Java. Полнотекстовый поиск нашел бы все упоминания термина “Java”, но многие абзацы могли бы быть нерелевантными, кроме того, абзацы, в которых нет термина “Java”, но относящиеся к языку программирования Java, не были бы найдены.

Допустим, что по запросу “Java” найдено много абзацев, и все они одинаково похожи на запрос. Как можно ранжировать такую поисковую выдачу? Для этого может быть использована функция семантики. Абзацы, которые лучше передают смысл документа, имеют большую релевантность.

Пусть к некоторым документам вручную привязан L-тег “Java” и определено значение функции семантики. В этом случае L-тег “Java” может участвовать в поиске вместе с другими L-тегами. Привязка поисковых запросов к документам вручную позволяет улучшить качество поиска в наиболее важных темах, кроме того, такой подход используется в рекламных системах.

Представленная модель позволяет вынести сложные вычисления оценки функции семантики на этап индексации, что снижает нагрузку на сервер в момент поиска. Кроме того, появляется возможность контролировать объем поискового индекса и, как следствие, нагрузку на сервер в момент выполнения поискового запроса. Это возможно за счет ограничения количества тегов по значению функции семантики.

## 6 Заключение

В работе представлена модель семантического поиска и продемонстрирована полезность тезаурусов типа WordNet. Дан небольшой обзор по типам тезаурусов и предложено решение некоторых проблем.

Были формализованы определения контекстов: понятия, абзаца, документа, запроса и пользователя. Были описаны алгоритмы для выделения контекстов с использованием большого корпуса текстов, наиболее полного тезауруса. Была уточнена модель семантического поиска, введенная ранее. Предложены способы оценки функций семантики и схожести с помощью различных контекстов, связей понятий из тезауруса. Была введена, но недостаточно формализована, функция близости понятий. Предполагается ее формализация в дальнейших работах. Кроме того, планируется:

- Описать особенности индексирования математических текстов.
- Рассказать о программной архитектуре, основанной на представленной модели.
- Оценить качество и быстродействие системы поиска по сравнению с другими решениями.

## Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект 17-07-00214).

## Литература

[1] Fellbaum, C.: WordNet. Blackwell Publishing Ltd, (1998)

[2] Malakhov, D., Sidorenko, Y., Ataeva, O., Serebryakov, V.: Semantic Search in a Personal Digital Library. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds). Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, 706. Springer, Cham (2017)

[3] Magnini, B., Strapparava, C.: Experiments in Word Domain Disambiguation For Parallel Texts. Proc. of the ACL-2000 Workshop on Word Senses and Multi-linguality. Association for Computational Linguistics, pp. 27-33 (2000)

[4] Miller, G.A., Fellbaum, C., Tengi, R.: WordNet. Cambridge, Princeton University (2006)

[5] Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval (1986)

[6] Лукашевич, Н.В.: Тезаурусы в задачах информационного поиска, М.: Изд-во МГУ (2011)

[7] Серебряков, В.А. Что такое семантическая цифровая библиотека In: RCDL 2014. сс. 21-25 (2014)