УДК 004.02

Формализация задачи семантического поиска

Д.А. Малахов
Московский государственный университет им. М.В. Ломоносова

Поиск информации является ключевой задачей в веке информационных технологий. Потребность в качественном удовлетворении информационной потребности человека растет так же быстро, как и объемы данных, с которыми приходится сталкиваться. Было проведено немало исследований для повышения качества поиска, были разработаны различные поисковые системы, но до сих пор задача является актуальной и ее решением занимаются ученые и инженеры по всему миру.

Одними из первых решений задачи поиска стали методы полнотекстового поиска, использующие только синтаксическую структуру текста, игнорируя смысл отдельных частей текста. Проект Semantic Web [1] был создан, чтобы исследовать возможности использования не только синтаксиса документа, но и семантики. В рамках проекта были разработаны такие стандарты, как язык OWL онтологий, формат представления знаний RDF, язык семантических запросов SPARQL и другие. К сожалению, несмотря на проделанную работу, на данный момент в прикладных системах, как правило, используется полнотекстовый поиск. Видимо, технологии, разработанные в рамках проекта Semantic Web не получили пока достаточного распространения из-за сложности построения поисковых запросов SPARQL и представления данных в формате RDF [2].

Несмотря на отсутствие явного прогресса в создании востребованной системы семантического поиска, стандарты, разработанные проектом Semantic Web, стали основой для исследований в смежных областях. Одной из таких областей является семантическое аннотирование, которое на данный момент активно исследуется и развивается [3][4]. Семантическое аннотирование позволяет привязывать к разным частям текста, понятия из онтологии. В результате процесса семантического аннотирования можно получить модель документа в терминах заданной онтологии. Результаты семантического аннотирования могут быть использованы в системе полнотекстового поиска для повышения качества поиска. Таким образом, на данный момент исследования направлены в сторону поиска гибридного решения, использующего как компоненты семантического поиска, так и компоненты полнотекстового поиска.

Данная работа посвящена формализации задачи семантического поиска. Семантическим поиском, как правило, называется процесс поиска документов по их содержанию. Нетрудно увидеть, что понятие семантического поиска недостаточно формально определено [5]. В частности, понятие содержания или смысла является многозначным. Далее будет предложена модель семантического поиска, основанная на использовании L-тегов. Модель L-тега является частным случаем модели S-тегов[6].

Пусть задан алфавит A, где алфавитом является любое конечное непустое множество.

Пусть задано множество терминов T, где термин является упорядоченным мультимножеством элементов из A.

Пусть задано множество L-тегов LT, где L-тег является упорядоченным мультимножеством элементов из T.

Пусть задано некоторое множество объектов O. Можно определить функции семантики и похожести:

$$semantic\colon (O, LT) \to R \quad (1)$$

$$similarity\colon (LT, LT) \to R \quad (2)$$

Функция семантики (1) оценивает то, насколько смысл терминов L-тега отражает смысл объекта, например, текстового документа. Функция схожести (2) оценивает то, насколько смысл первого L-тег обобщает смысл второго L-тега. Чем более общим является первый L-тег по отношению ко второму L-тегу, тем меньше функция схожести (2).

Пусть функция семантики (1) известна в некоторых точках P.

Пусть функция схожести известна всюду.

Тогда мы можем интерполировать функцию семантики (1):

$$\begin{cases} semantic(O, LT),\ (O, LT) \in P \\ \max\limits_{(O, LT1) \in P} (semantic(O, LT1) * similarity\ (LT1,\ LT)), (O, LT) \notin P \end{cases} \quad (3)$$

Модель семантического поиска по объектам из множества O:

- Пусть поисковой запрос Q представляет собой L-тег.
- Пусть задана функция семантики (1) для некоторых пар $(o, lt) \in P$.
- Пусть задана функция схожести (2) между L-тегами из LT.
- Результатом исполнения запроса Q является множество L-тегов, с привязанными к ним функцией семантики (1) документами, отсортированное по значениям интерполированной функции семантики (3) для пар (o, Q), где $o \in O$.

Использование представленной модели позволяет свести задачу семантического поиска к решению задачи определения функции схожести (2) между небольшими текстами и задаче определения функции семантики (1) между небольшим текстом и объектом/документом. Это позволяет заранее выделить/присвоить отражающие смысл объекта L-теги, рассчитав функцию семантики (1), и использовать результаты при поиске.

В качестве интерпретации модели семантического поиска рассмотрим поиск по хэштегам.

Любой хэштег является термином и L-тегом, образуя множество LT. Хэштеги привязываются к документам из множества O. Пусть заданы пары документов и привязанных к ним хэштегов P. Можно определить функцию схожести (2) и семантики (1):

$$similarity\ (LT1,\ LT1) = \begin{cases} 0, & LT1 \neq LT2 \\ 1, & LT1 = LT2 \end{cases}$$

$$semantic(O,\ LT) = \begin{cases} 0, & (O, LT) \in P \\ 1, & (O, LT) \notin P \end{cases}$$

Модель хэштега достаточно тривиальна и не позволяет использовать все возможности функции схожести и семантики, но позволяет продемонстрировать универсальность модели.

Можно рассмотреть более сложные интерпретации модели L-тега:

- Словосочетание;
- Предложение;
- Абзац.

В этом случае поисковой запрос на естественном языке также является L-тегом.

Функция схожести (2) можно определить как релевантность первого L-тега ко второму L-тегу. Таким образом, для оценки функции схожести (2) можно использовать различные алгоритмы определения релевантности.

Функция семантики (1) определяется через выделение/привязку L-тегов. При ручной привязке L-тегов значение функции семантики (1) задает человек. Например, автор описывает ключевые слова статьи и оценивает их значимость. Для автоматического выделения L-тегов характерно:

- Получение кандидатов в L-теги. Кандидаты могут быть получены из самого текста, так и из заданного набора, например, тезауруса.
- Вычисление функции семантики (1). Может использоваться информация о контексте и тезаурусы.
- Фильтрация кандидатов. Может быть задано некоторое ограничение на количество L-тегов или на минимальное значение функции семантики (1).

При выборе интерпретации модели L-нужно понимать что, чем больше терминов содержит L-тег, тем больше смысла он в себе несет. Чем больше смысла в L-теге, тем большему количеству запросов он будет удовлетворять. От этого появляется больше потенциальных ложноположительных результатов. С другой стороны, увеличивается полнота поиска. Не существует универсальной интерпретации L-тега. Интерпретации модели L-тега должна быть подобрана под конкретную задачу.

Литература

1. *Berners-Lee T. et al.* The semantic web // Scientific american.-2001.-Т. 284.-С. 28-37.
2. *Малахов Д.А.* Проблемы семантического поиска // Труды 58-й научной конференции МФТИ. 2015.
3. *Alahmari F., Magee L.* Linked Data and Entity Search: A Brief History and Some Ways Ahead //Australasian Web Conference 2015. – 2015. – Т. 27. – С. 29.
4. *Berlanga R., Nebot V., Pérez M.* Tailored semantic annotation for semantic search //Web Semantics: Science, Services and Agents on the World Wide Web. – 2015. – Т. 30. – С. 69-81.
5. *Серебряков В. А.* Что такое семантическая цифровая библиотека // RCDL. – 2014. – С. 21-25.
6. *Малахов Д.А., Сидоренко Ю.А., Атаева О.М., Серебряков В.А.* Что такое семантическая цифровая библиотека // DAMDID/RCDL. – 2016. – С. 148-155.