Исследование качества автоматической классификации текстовых документов с использованием семантического графа документа

Analyzing the Quality of Automated Indexing of Text Documents Using Their Semantic Grafs

Дослідження якості автоматичної класифікації текстових документів з використанням семантичного графу документів

Соколовский В. В.
Государственная публичная научно-техническая библиотека России, Москва, Россия

Vladimir V. Sokolovsky
Russian State Pubic Library for science and Technology, Moscow, Russia

Соколовський В. В.
Державна публічна науково-технічна бібліотека Росії, Москва, Росія

Проведено сравнение классификации с использованием двух разных подходов к мере подобия текстовых документов. В одном случае документы считались в той мере подобными друг другу, в какой подобен состав входящих в них ключевых слов. В другом случае документы считались подобными друг другу в той мере, в какой подобны их семантические графы.

The accuracy of indexing is compared between the two different approaches towards the of similarity text documents. In one case, the documents are regarded as similar to the degree of the similarity of the keywords they comprise. In another, the documents are considered similar to the degree of similarity of their semantic graphs.

Проведено порівняння класифікації з використанням двох різних підходів до міри схожості текстових документів. В одному випадку документи вважаються схожими в тій мірі, в якій є схожим склад ключових слів, що входять до них. В іншому випадку документи вважаються схожими в тій мірі, в якій є схожими їхні семантичні графи.

Под классификацией понимают отнесение классифицируемых объектов к заранее определённым категориям. Для целей информационного поиска обычно различают два типа классификаций. Первый тип — классификация терминов, целью которой является группировка терминов в классы. Классификация терминов может использоваться при необходимости для обогащения запроса низкочастотными терминами, аналогично тезаурусу. И второй тип это классификация документов, целью которой является улучшение результативности и оперативности поиска за счёт обращения только к определённым частям массива документов, среди которых производится поиск.

Для классификации (а также, для кластеризации) документов вводится мера подобия (близости, сходства) документов между собой. Обычно текстовые документы считаются в той мере подобными друг другу, в какой подобен их терминологический состав.

Между тем, текстовые документы на естественном языке можно формализовать в другом виде. Давно известной и удобной моделью представления информации является представление в виде графа, в узлах которого находятся объекты, а рёбра графа представляют собой отношения (из некоторого словаря отношений) между этими объектами. Примером активного применения такой графовой модели для хранения информации в настоящее время является RDF (Resource Description Framework) — модель описания ресурсов, предназначенная для хранения Web метаданных (информации о Web ресурсах и системах их использующих, описания содержимого, возможностей и т. д.). Основные цели использования графовой модели — сделать машинно-обрабатываемой представленную в таком виде информацию. Естественно, такое метаописание Web ресурсов кто-то должен создать, то есть такие данные сразу создаются человеком или компьютером в графовом виде.

Другой подход, связанный с автоматическим преобразованием текста на естественном языке к формальному графовому виду (рис.1), практикует рабочая группа Aot.ru. Согласно этому подходу, грамотная декомпиляция языковых механизмов позволит максимально приблизить человеческий язык к современному компьютеру. То есть, цель та же — сделать информацию машинно-обрабатываемой. Рабочая группа Aot.ru разрабатывает программное обеспечение в области автоматической обработки текста, в основном связанное с анализом русского языка. Один из их проектов, семантический анализ текста на русском языке, представляет собой построение семантического графа текста. Семантический анализ строит семантическую структуру одного предложения на русском языке. Семантическая структура состоит из семантических узлов и семантических отношений.

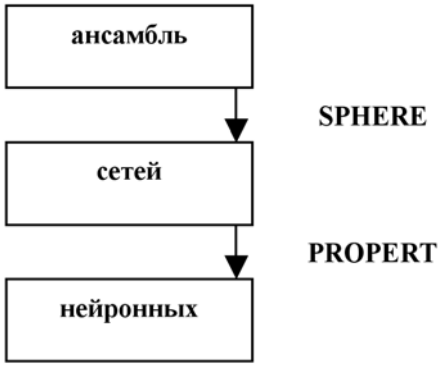

Рис.1. Представление предложения *«Ансамбль нейронных сетей»* в виде семантического графа

Используя эту модель представления информации в виде графа, было проведено сравнение точности классификации с использованием двух разных подходов к мере подобия текстовых документов. В одном случае документы считались в той мере подобными друг другу, в какой подобен состав входящих в них ключевых слов. В другом случае документы считались подобными друг другу в той мере, в какой подобны их семантические графы. Для преобразования текстовых документов на русском языке к виду семантического графа были использованы программные библиотеки, предоставленные Aot.ru в бесплатный доступ с лицензией LGPL.

В качестве материала для эксперимента по сравнению точности классификации были выбраны два узкотематических набора документов. Идея эксперимента в том, что документы подвергаются автоматической классификации, результаты которой сравниваются с заведомо известной классификацией — вхождением документа в одну из двух групп. Одна группа документов — рефераты на статьи журнала «Artificial Intelligence» за 2002 год (всего выбрано 57 статей, назовём эту группу **G0**). Другая группа документов — рефераты на статьи журнала «C Users Journal», а также «C/C++ Users Journal» за 1991-1997 годы (всего выбрано 58 статей, назовём эту группу **G1**). Рефераты были взяты из базы данных реферативного журнала «КОМПЬЮТЕРНЫЙ ВЕСТНИК».

Эксперимент проводился в следующем порядке. Строилась таблица, в которой были вычислены меры подобия каждого документа с каждым. Документы считались похожими, если мера подобия превышала пороговое значение подобия **P**. Затем из участия в автоматической классификации были исключены документы, которые были подобны менее чем некоторому заданному количеству **N** документов.

Для начала пороговое значение **N** было подобрано достаточно большим, чтобы результаты автоматической классификации получились наиболее хорошими. Таким образом, если в начале эксперимента было взято 57 и 58 статей, то после такого исключения остались 17 и 5 статей соответственно. Остальные были исключены как непригодные для классификации по той причине, что нашлось слишком мало (меньше выбранного числа **N**) похожих на них статей, чтобы можно было делать выводы о принадлежности к той или иной группе. В этом абзаце речь идёт о способе вычисления меры подобия документов по семантическому графу, поскольку при вычислении меры подобия документов по ключевым словам, в необходимой мере похожих (больше чем **P**) документов всегда находилось достаточно (больше выбранного числа **N**).

Итак, для каждого документа **Дi** были найдены похожие на него документы **{Д}i** (часть которых принадлежит к одной, а оставшаяся часть к другой узкоспециальной группе, **G0** и **G1** соответственно). После этого меры подобия найденных документов **{Д}i** суммировались (отдельно для одной и другой группы). И текст автоматически классифицировался как принадлежащий к той группе, для которой сумма мер подобия найденных документов **{Д}i** (похожих на данный документ **Дi**) оказалась больше.

Для того случая, когда **N**, было выбрано большим и на классификацию осталось 17 и 5 документов соответственно, все они оказались классифицированы правильно (Таблица 1.). То есть 17 из документов попали в свою группу **G0**, а остальные 5 попали в свою группу **G1**, для обоих способов вычисления меры подобия.

После того как пороговое значение **N** было уменьшено, на классификацию попало 24 и 28 документов соответственно. В этом случае количество ошибочно классифицированных текстов составило половину (для семантического графа) и более (для ключевых слов) документов. Причём подавляющее число ошибок пришлось на тексты, относящиеся к группе **G1**, а группа **G0** содержит приемлемое количество ошибок. Объясняется такая ситуация по всей видимости тем, что **G0** представляет из себя действительно группу однородных документов, в то время как **G1** таковой не является ни по составу ключевых слов, ни по семантическим графам документов.

Оказалось, что документы из группы **G1**, которые были правильно классифицированы при малом **N**, для меры подобия, вычисляемой из семантических графов, оказались теми же, что были правильно классифицированы в эксперименте с достаточно большим **N**. Для меры подобия, вычисляемой по ключевым словам, такого эффекта не было. Следовательно, при применении меры подобия, вычисляемой из семантического графа, результат классификации получается более устойчивый к изменению порогового значения **N**. И связано это с большей точностью и меньшей полнотой указанного метода поиска похожих документов. Когда полнота метода поиска похожих документов по ключевым словам уже даёт слишком большой шумовой вклад, то метод поиска по семантическим графам работает лучше.

Таблица 1

Результаты классификации

N	**Кол-во док-в переданных на классификацию G0/ G1**	**Правильно классифицир. G0/ G1**	**Неправильно классифицир. G0/ G1**	**Мера подобия**
достаточно большое	17/5	17/5	0/0	оба способа
уменьшено	24/28	19/0	5/28	ключев. слова
уменьшено	24/28	22/5	2/23	семант. граф

Таким образом, классификация текстовых документов на основе их графового представления оказалась более точной и менее полной, чего и следовало ожидать. В качестве дальнейшего направления исследования этой темы видится интересным провести аналогичное исследование для кластеризации документов.

Литература

1. Brian McBride, Resource Description Framework (RDF): Concepts and Abstract Syntax. // http://www.w3.org/TR/rdf-concepts/
2. Автоматическая Обработка Текста. Технологии. // http://www.aot.ru/technology.html
3. Автоматическая Обработка Текста. Dialing Graph Builder. // http://www.aot.ru/cgi-bin/translate.exe?graph_action