# ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ТЕМАТИЧЕСКОГО КОРПУСА ВОПРОСОВ И ОТВЕТОВ

Белобородов А.В., Кузнецов А.М.
*e-mail: whoisnexta@gmail.com*

Вопросно-ответный сервис (community question answering, CQA) — это система, позволяющая обычным пользователям сети Интернет задавать вопросы в свободной форме или отвечать на них (более подробное описание дается в [1]). Некоторые из таких систем имеют тематическую направленность, чтобы привлекать к общению пользователей с общими интересами (например, stackoverflow.com — CQA-ресурс, посвященный программированию). Наличие тематики позволяет извлекать из подобных сервисов информацию более качественно и детализированно, отталкиваясь от специфичных терминов.

Была поставлена задача: из набора вопросов и ответов про медицину извлечь информацию о болезнях, которыми интересуются пользователи, и о медикаментах, которые пользователи советуют для лечения. Эксперимент проводился на корпусе вопросов и ответов сервиса Ответы@Mail.Ru[1]. Были взяты данные категорий «Болезни, Лекарства», «Врачи, Клиники, Страхование», «Детское здоровье», «Отвечает врач» за период с 01.04.2011 по 31.03.2012 — всего 128370 вопросов с ответами. Среднее количество ответов на вопрос — 5, средняя длина вопроса в словах — 7. Все слова в тексте вопросов были лемматизированы, вопросы и ответы были очищены от HTML разметки, знаков препинания.

Для идентификации терминов в тексте был составлен список из 1049 заболеваний, собранных на основе справочника практического врача [3] и список медикаментов из государственного реестра лекарственных средств[2](11926 уникальных наименований по состоянию на сентябрь 2012 года). Был разработан метод автоматического извлечения терминов из корпуса вопросов и ответов. Так как названия заболеваний и лекарств носят профессиональный характер, была выдвинута гипотеза, что в них часто допускаются орфографи-

[1]http://otvet.mail.ru
[2]http://grls.rosminzdrav.ru/grls.aspx

ческие ошибки. На основе работы [2] был реализован нечеткий поиск с помощью индекса 3-грамм и последующего вычисления расстояния Левенштейна в зависимости от длины слова. Например, для слова диабет (723 вхождения) были найдены слова диабед (16), диобет (5), деабет (3), диабетя (1), диает (1), диабеа (1), 7диабет (1). Исправление опечаток позволило найти дополнительные вхождения болезней и медикаментов: добавилось 1110 вхождений заболеваний (46781 было без опечаток) и 5456 вхождений медикаментов (248954 было). Из всех вопросов выбирались удовлетворяющие следующим условиям:

1. В тексте вопроса упоминается заболевание из составленного списка

2. В тексте ответов на вопрос упоминается медикамент из составленного списка

Тем самым моделировалась ситуация, когда от спрашивающего пользователя поступает жалоба на заболевание, а от отвечающих пользователей — совет использовать определенное лекарство. Всего найдено 24972 таких вопроса (19%). Далее пары (заболевание, лекарство) упорядочивались по количеству вопросов, в которых были встречены. Очевидно, что если пара встретилась в небольшом количестве вопросов, о ней нельзя сказать ничего определенного — это может быть несколько случайных упоминаний не по теме. Напротив, наиболее частые пары выражают мнение людей о конкретном лекарстве относительно конкретной болезни.

Таблица 1: Наиболее часто упоминаемые медикаменты для некоторых заболеваний

| Ангина | | Рана | | Герпес | |
|---|---|---|---|---|---|
| Люголь | 134 | Перекись | 149 | Ацикловир | 286 |
| Ромашка | 125 | Зеленка | 96 | Зовиракс | 136 |
| Йод | 123 | Йод | 88 | Сера | 92 |
| Фурацилин | 105 | Левомеколь | 73 | Корвалол | 51 |
| Шалфей | 86 | Спирт | 51 | Фенистил | 34 |

Полные данные таблицы доступны в сети Интернет[3]. В будущих исследованиях планируется оценить точность извлекаемой информации о заболеваниях и медикаментах с помощью международной классификации болезней МКБ-10, которая находится в открытом доступе в сети Интернет[4]. В МКБ-10 для большинства заболеваний представлен список препаратов, рекомендуемых для лечения. Если точность извлекаемых данных окажется на достаточно высоком уровне, то рассмотренный метод можно будет использовать для автоматического извлечения тематических данных из содержимого вопросно-ответных сервисов.

**Литература**


[1] *Белобородов А.В.* Сравнение трех мер семантической близости вопросов в социальных вопросно-ответных сервисах // Труды Международной (43-й Всероссийской) молодежной школы-конференции «Современные проблемы математики», 2012.

[2] *Norvig P.* How to Write a Spelling Corrector // Online; visited February 22, 2008. http://norvig.com/spell-correct.html.

[3] *Воробьев А.И.* Справочник практического врача // М.: Медицина, 1981, 656 с.


[3] https://raw.github.com/bellal89/cqa_medical/Sopromat2013/src/Sopromat2013/deseases-medicaments.txt

[4] http://www.rlsnet.ru/mkb_tree.htm