ИЗВЛЕЧЕНИЕ СУЩНОСТЕЙ ИЗ ТЕКСТОВ В СИСТЕМЕ КОНТЕНТ-МОНИТОРИНГА ИНТЕРНЕТ-РЕСУРСОВ

***Ландэ** Дмитрий Владимирович, **Дармохвал** Александр Теодорович*

Информационный центр «ЭЛВИСТИ»
Киев, Украина
dwl@visti.net, hval@visti.net

Извлечение сущностей [1] является важным этапом автоматизированной обработки текстов в таких приложениях, как информационно-поисковые и информационно-аналитические системы, системы поиска плагиата, машинного перевода и т.д. В качестве сущностей в рамках доклада рассматриваются названия компаний.

При выделении названий компаний авторами используется метод маркеров и шаблонов. В качестве маркеров используются «префиксы» компаний (ООО, АОЗТ, ЗАО и т.п.), «суффиксы» для зарубежных компаний (например, Corp., Ltd., Inc. и т.д.), должности персонала, имена. В качестве шаблонов используются варианты написания известных компаний и общепринятые правила, используемые в новостных сообщениях (например, кириллические названия компаний берутся в кавычки, а латинские – не берутся, названия начинаются с прописных букв или цифр и т.д.). Для реализации данного подхода использовались средства регулярных выражений языка программирования Perl. Вместе с тем требования к быстродействию обуславливают необходимость распараллеливания процессов извлечения сущностей и использования языков программирования более низкого уровня.

Описанный в докладе подход к извлечению сущностей нашел применение в системе контент-мониторинга InfoStream, входной поток которой составляет до 100 тысяч документов в сутки [2]. Подход оказался эффективным для новостных сообщений, представленных на украинском, русском и английском языках. Полнота извлечения названий компаний составляет около 85 %, точность – 98 %.

ЛИТЕРАТУРА

[1] Rosenfeld B., Feldman R., Fresko M., TEG – A Hybrid Approach to Information Extraction // Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004. – pp. 589–596.

[2] Григорьев А.Н., Ландэ Д.В. Система мониторинга новостей InfoStream - информационное пространство из одних рук // Построение информационного общества: ресурсы и технологии. Тезисы докладов и информационные материалы XI международной научно-практической конференции. – К: УкрИНТЭИ, 2005. – С. 17–20.