

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет радиофизики и компьютерных технологий
Кафедра интеллектуальных систем

Аннотация к дипломной работе

**«Алгоритм извлечения фактов из опубликованных
вакансий в интернете»**

Куксенюк Тимофей Сергеевич

Научный руководитель: старший преподаватель, Адуцкевич И. А.

РЕФЕРАТ

Дипломная работа: 48 страниц, 9 рисунков, 2 таблицы, 41 источник.

АЛГОРИТМЫ ФИЛЬТРАЦИИ ИНФОРМАЦИИ, ИЗВЛЕЧЕНИЕ ФАКТОВ, СЕМАНТИЧЕСКИЙ АНАЛИЗ, ИЗВЛЕЧЕНИЕ ТЕКСТА ИЗ ВЕБ-СТРАНИЦ.

Объект исследования: опубликованные вакансии в интернете.

Предмет исследования: алгоритм извлечения фактов из опубликованных вакансий.

Целью дипломной работы является изучение и разработка алгоритма извлечения фактов из текстов опубликованных вакансий, который поможет человеку найти нужную вакансию.

Методами исследования являлись теоретический анализ литературы по теме работы и эмпирический анализ эффективности существующих методов извлечения фактов, а также методов фильтрации информации на странице.

В настоящей работе рассмотрены системы и методы фильтрации информации на странице, проанализированы сторонние алгоритмы, методы, технологии и сервисы для извлечения фактов из текстов, изучены этапы извлечения фактов.

В результате эмпирического анализа существующих алгоритмов фильтрации информации и методов извлечения фактов из текста выявлены основные сложности и недостатки этих алгоритмов, предложены методы их решения. А также мной была разработана структура базы данных вакансий и собственный алгоритм для фильтрации информации с веб-страницы и поиска фактов из полученного текста.

РЭФЕРАТ

Дыпломная праца: 48старонак, 9малюнкаў, 2 табліцы, 41 крыніца.

АЛГАРЫТМЫ ФІЛЬТРАЦЫІ ІНФАРМАЦЫІ, ВЫМАННЕ ФАКТАЎ, СЕМАНТЫЧНЫ АНАЛІЗ, ВЫМАННЕ ТЭКСТУ З ВЭБ-СТАРОНАК.

Аб'ект даследавання: апублікаваныя вакансіі ў інтэрнэце.

Прадмет даследавання: алгарытм здабывання фактаў з апублікаваных вакансій.

Мэта дыпломнай працы: вывучэнне і распрацоўка алгарытму здабывання фактаў з тэкстаў апублікаваных вакансій, які дапаможа чалавеку знайсці патрэбную вакансію.

Метадамі даследавання з'яўляліся тэарэтычны аналіз літаратуры па тэме работы і эмпірычны аналіз эфектыўнасці існуючых метадаў здабывання фактаў, а таксама метадаў фільтравання інфармацыі на старонцы.

У працыразгледжаны сістэмы і метады фільтравання інфармацыі на старонцы, прааналізаваныя іншыя алгарытмы, метады, тэхналогіі і сэрвісы для здабывання фактаў з тэкстаў, вывучаны этапы здабывання фактаў.

У выніку эмпірычнага аналізу існуючых алгарытмаў фільтрацыі інфармацыі і метадаў здабывання фактаў з тэксту выяўлены асноўныя складанасці і недахопы гэтых алгарытмаў, прапанаваны метады іх рашэння. Таксама мной была распрацавана структура базы дадзеных вакансій і ўласны алгарытм для фільтрацыі інфармацыі з вэб-старонкі і пошуку фактаў з атрыманага тэксту.

ABSTRACT

Thesis: 48 pages, 9 figures, 2 tables, 41 sources.

ALGORITHMS OF INFORMATION FILTRATION, EXTRACTION OF FACTS, SEMANTIC ANALYSIS, EXTRACTION OF TEXT FROM WEB-PAGES.

Object of research: published vacancies on the Internet.

Subject of research: algorithm for extracting facts from published vacancies.

The purpose of research: studying and developing an algorithm for extracting facts from the texts of published vacancies that will help a person find a suitable vacancy.

Research methods are theoretical analyze of literature about the problem and empirical analysis of the effectiveness of existing methods of extracting facts, and information filtering methods on a page.

In the present work, systems and methods of filtering information on the page are examined; algorithms, methods, technologies and services to extract facts from texts are analyzed, stages of extracting facts were studied.

As a result of the empirical analysis of existing information filtering algorithms and methods of extracting facts from the text, the main difficulties and disadvantages of these algorithms have been revealed, and methods for their solution have been proposed. I also developed the structure of the database of vacancies and my own algorithm for filtering information from a web page and finding facts from the resulting text.