Восстановление связей между библиографическими записями

Князева А.А., Колобов О.С.

Институт вычислительных технологий СО РАН
Институт сильноточной электроники СО РАН

Определение понятия связывания

Связывание записей - сравнение информации из различных источников данных с целью определения, какие пары записей представляют один и тот же объект реального мира.

Частный случай - выявление дубликатов.

Постановка задачи

Автоматически связывать библиографическую и авторитетную запись, если они относятся к одному автору.

Библиографическая запись

- элемент библиографической информации, фиксирующий в документальной форме сведения о документе, позволяющие его идентифицировать, раскрыть его состав и содержание в целях библиографического поиска (ГОСТ 7.76).

Авторитетная / Нормативная запись

- машиночитаемая запись, исходным элементом которой является принятый заголовок. Кроме того, авторитетная / нормативная запись может содержать параллельные заголовки, ссылки, справки, примечания каталогизатора, сведения об источниках информации и о библиографирующем учреждении, ответственном за запись, а также другую информацию, характеризующую принятый заголовок (RUSMARC).

Обзор

Первое упоминание задачи связывания — Newcombe (1959 г.)

Формальная математическая модель — I.P. Fellegi and A.B. Sunter (1969 г.)

В настоящее время на этой идее основано целое семейство вероятностных моделей.

Существующие системы связывания записей

MARLIN (Bilenko, M; Mooney, RJ. - The University of Texas at Austin);

TAILOR (Elfeky, M. - Drexel University, Philadelphia);

Febrl (P. Christen and T. Churches. - Australian National University, Canberra, Australia);

VIAF (IFLA).

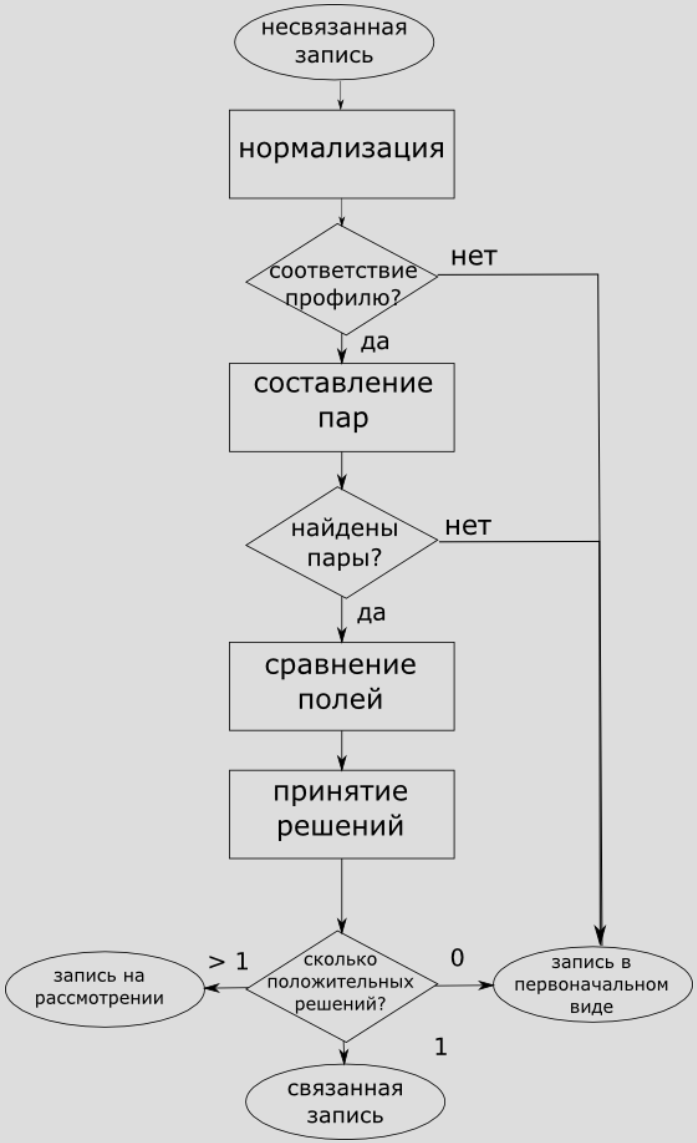
Отличия предлагаемого подхода

Работа с записями в формате RUSMARC;

Связывание записей разных типов;

Обучающая и тестовые выборки.

Модель процесса связывания



Дополнительные модули:

- Обучение решающей функции;
- Оценка качества связывания.

Блок нормализации

Назначение блока:

- Очистка и стандартизация значений отдельных полей;
- **Проверка на соответствие профилю.**

Блок составления пар

Назначение блока - сокращение перебора и нахождение записей-кандидатов для связывания.

Методы составления пар:

- **Стандартных блоков;**
- Ближайших соседей;
- Bigram-индексирования и др.

Расширенная авторитетная запись

Состоит из самой авторитетной записи и всех библиографических записей, находящихся в базе данных и связанных ней.

Сравнение с расширенной авторитетной записью позволяет увеличить количество информации для связывания.

Блок сравнения отдельных полей

Назначение блока - вычислить соответствие на уровне полей записей.

Методы сопоставления строк:

- Символьные;
- **На основе токенов;**
- На основе Q-грамм.

Блок вынесения решения

Назначение блока - принять решение о соответствии либо несоответствии записей друг другу.

Возможные методы:

- Набор эмпирических правил;
- **Индукционная модель;**
- Кластерная модель;
- Гибридная модель.

Эксперимент

Коллекции, предоставленные НП МедАрт:
1. Библиографическая БД, около 300000 зап.
2. Авторитетная БД имен лиц, около 10 000 зап.

Тестовая выборка состояла из 624 пар записей.

Классифицировано верно — 622 пары (99,7%)
Классифицировано ошибочно — 2 пары (0,3%)

Заключение

1. Модель автоматического связывания.

2. Особенности модели:
   Обучение;
   Расширенные авторитетные записи;
   Расширяемость.

Восстановление связей между библиографическими записями

Князева Анна

amili@mail.ru

Спасибо за внимание!