Автоматическое связывание документов

Князева А.А., Турчановский И.Ю., Колобов О.С.

Институт вычислительных технологий СО РАН
Институт сильноточной электроники СО РАН

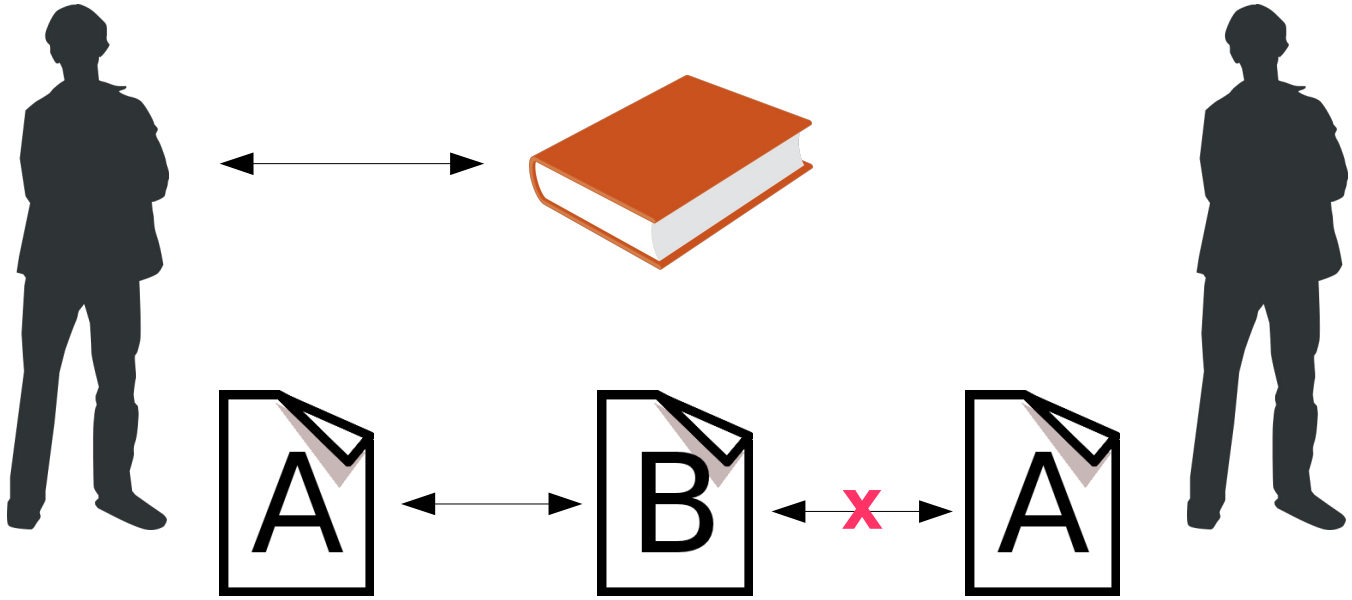
Определение понятия связывания

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

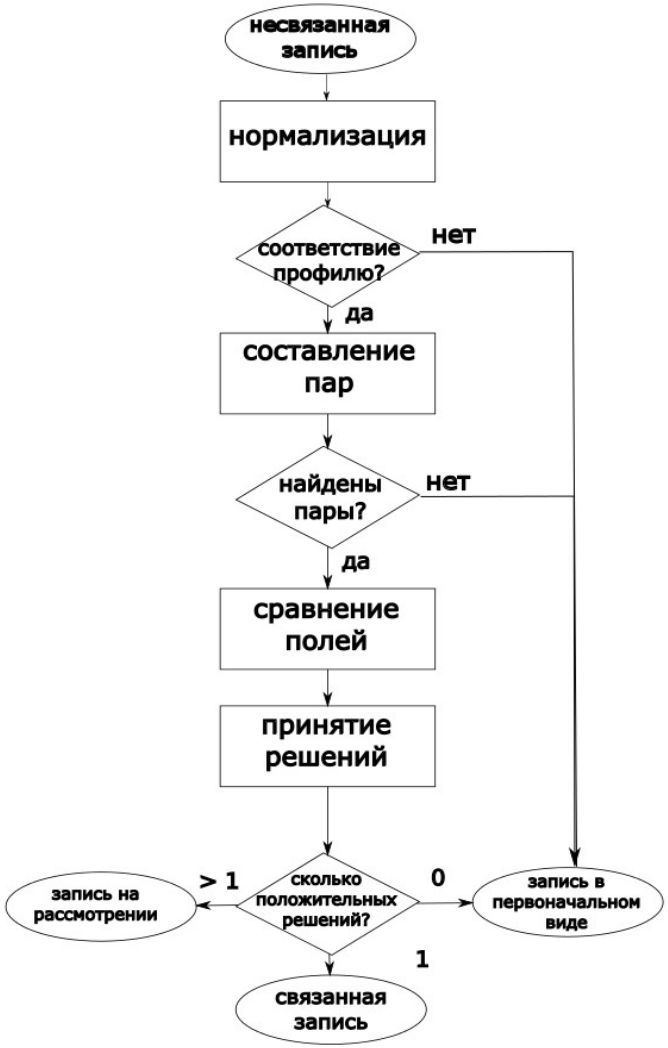
Частный случай - выявление дубликатов.

Постановка задачи

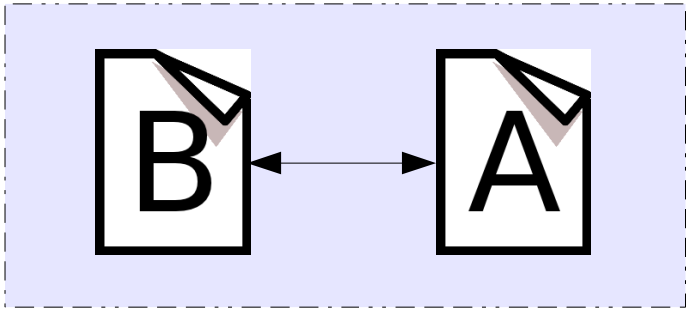
Автоматически связывать библиографический и авторитетный документы, если они относятся к одному автору.



Алгоритм процесса связывания



Составляем пары:

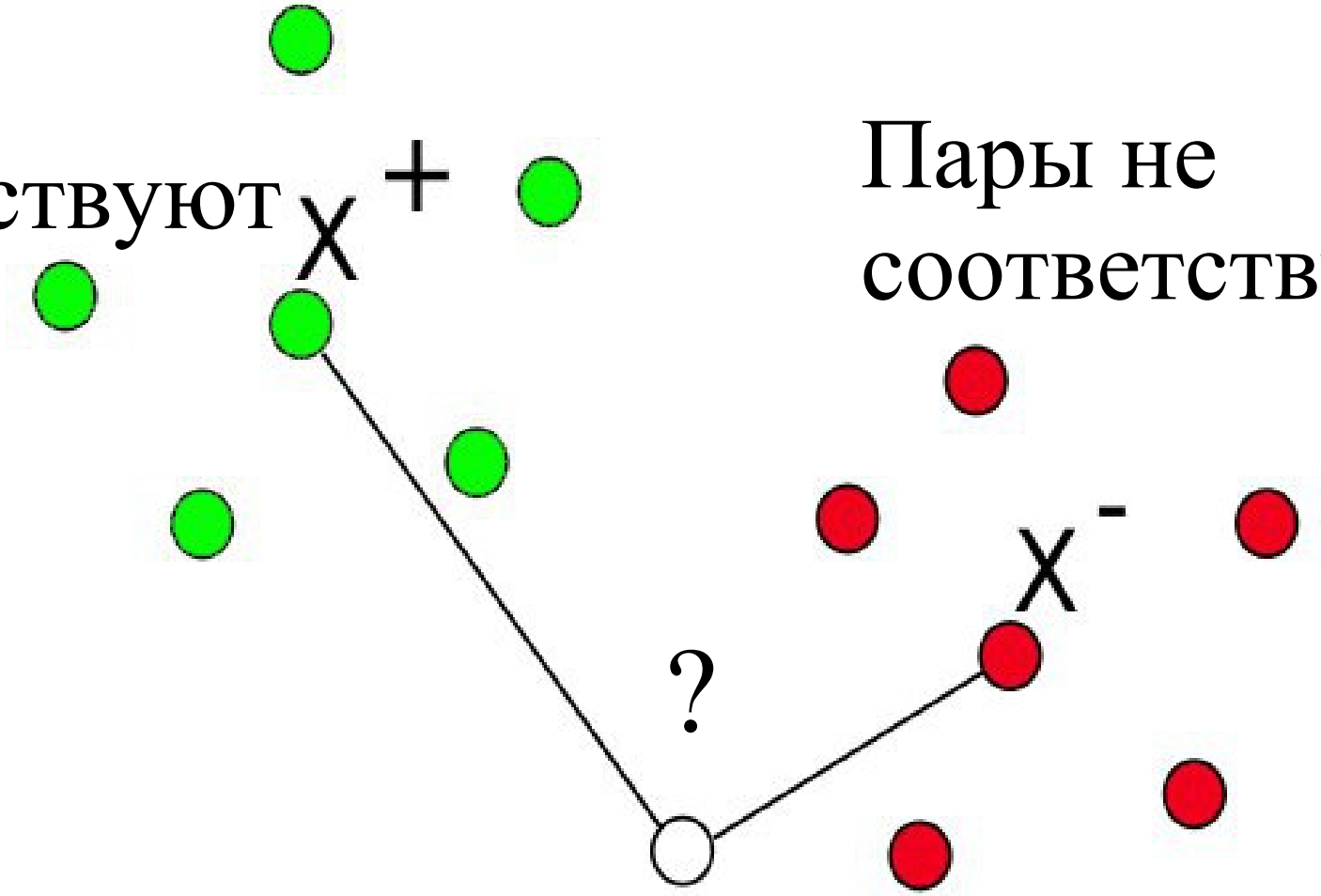


Вычисляем характеристики:

$$X=(X_1, \ldots X_k)^T$$

Принятие решения

Пары соответствуют



Пары не соответствуют

Эксперимент на реальных данных

Коллекции, предоставленные НП «МедАрт»:

- Библиографическая БД, около 300 000 записей;
- Авторитетная БД имен лиц, около 10 000 записей.

Список из 42 фамилий с инициалами, соответствующих сразу двум авторам.

1279 пар из авторитетной и библиографической записей.

Фрагмент библиографической записи

001 П15/А437-114799

701 1$аПанов $b А. А. $g Андрей Алексеевич $c физиология $f 19700224 $3APanov_AndrA200405076348070 0 $p кафедра нормальной физиологии ПФ СГМУ

701 1 $a Ковалев $b И. В. $p Сибирский медицинский университет (Томск)

701 1 $a Бородин $b Ю. Л.

71202 $a Сибирский медицинский университет $c Томск

606 1# $a МЫШЦА ГЛАДКАЯ $x действия лекарственных препаратов $3 D009130Q000187 $ 2mesh $8 rus

606 1 $a ФЕНИЛЭФРИН $x терапевтическое применение $3 D010656Q000627 $2 mesh $8 rus

Фрагмент авторитетной записи

001 APanov_AndrA2004050763480700

200 1$a Панов $b А. А. $c физиология $f 19700224 $g Андрей Алексеевич $f Томск

830 $a Образование: в 1995 г. окончил Сибирский медицинский университет (Томск), с 1995 по 1998 гг. - аспирант кафедры биофизики Сибирского медицинского университета (Томск).

$a Трудовая деятельность: с 1999 по 2001 гг. - ассистент кафедры нормальной физиологии Сибирского медицинского университета (Томск), с 2002 г. - старший преподаватель кафедры нормальной физиологии педиатрического факультета Сибирского медицинского университета...

Переменные основной группы

Переменная	А3	Б3
Out - соответствие	001	700$3
Birth - дата рождения	200$f	700$f
Death - дата смерти	200$f	700$f
Addition - профессия	200$c	700$c
Work1 - место работы автора	830$a	700$p
Work2 - коллектив	830$a	712$a
Place1 - географическое дополнение	200$y	712$c

Переменные основной группы
(рабочий пример)

Переменная	А3	Б3
Birth	1970	1970
Death	-	-
Addition	физиология	физиология
Work1	...преподаватель кафедры нормальной физиологии педиатрического факультета Сибирского медицинского университета...	кафедра нормальной физиологии ПФ СГМУ
Work2	Сибирского медицинского университета...	Сибирский медицинский университет
Place1	Томск	Томск

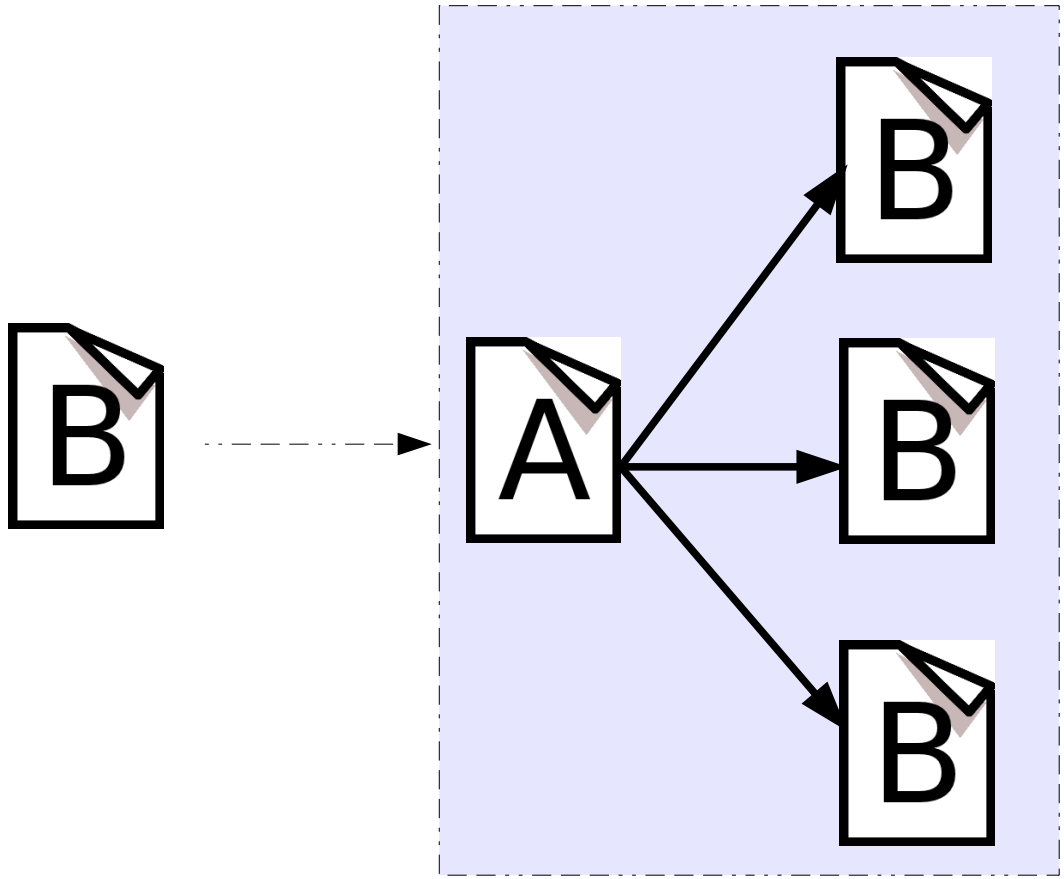
Этап 1. Только основная группа

(6 факторных переменных)

100 итераций

Требования к записям	Обучающая / тестовая выборка	Среднее кол-во ошибок в тесте	Средний процент ошибок в тесте
Присутствует 1 поле или больше (28%)	816 / 400	9,43	2,36%
Присутствует 2 поля или больше (21%)	581 / 400	2,49	0,62%

Расширенная авторитетная запись



Переменные доп. группы

Информация	Поля БЗ	Кол-во	Доля	Максимум БЗ
Соавторы (фамилия и инициалы)	700$a, 700$b	Coauthor1	Coauthor2	Coauthor3
Соавторы (код)	700$3	CoauthorId1	CoauthorId2	CoauthorId3
Предметные рубрики (название)	606$a	Subject1	Subject2	Subject3
Предметные рубрики (код MeSH)	606$3	SubjectId1	SubjectId2	SubjectId3

Вместе с основной группой 18 факторных переменных

Этап 2: Основная и дополнительная группа

(18 факторных переменных)

100 итераций

Требования к записям	Обучающая / тестовая выборка	Среднее кол-во ошибок в тесте	Средний процент ошибок в тесте
Присутствует 1 поле или больше (97,7%)	879 / 400	9,44	2,36%
Присутствует 2 поля или больше (77,3%)	875 / 400	8,59	2,15%

Расширенная группа переменных

- Основная группа;
- Дополнительная группа;
- Расширение основной группы.

22 факторные переменные

Переменная	БЗ
Addition2 - профессия	700$c
Work3 - место работы автора	700$p
Work4 - коллектив	712$a
Place2 - географическое дополнение	712$c

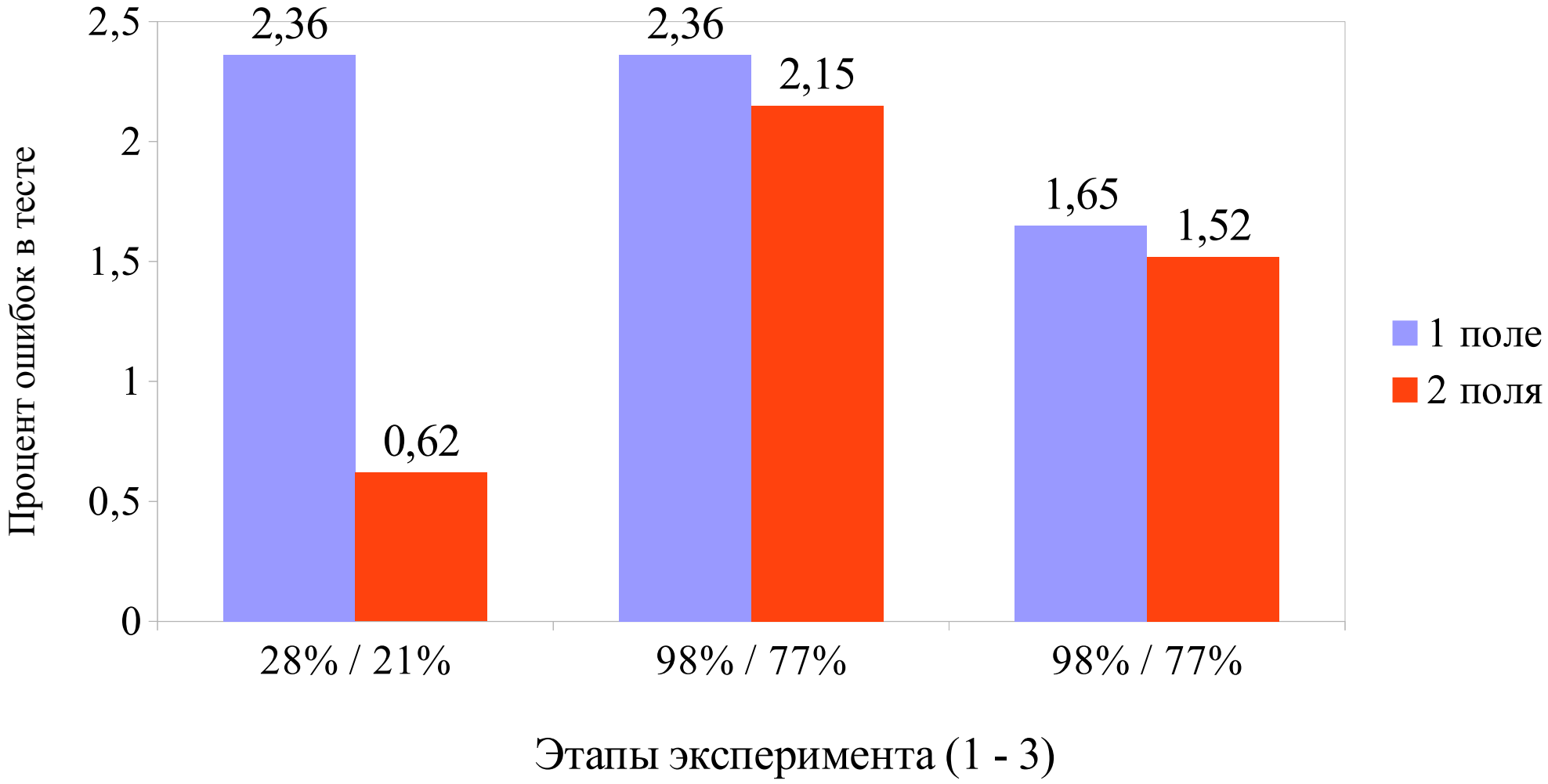
Этап 3: Расширенная группа

(22 факторные переменные)

100 итераций

Требования к записям	Обучающая / тестовая выборка	Среднее кол-во ошибок	Средний процент ошибок
Присутствует 1 поле или больше (97,7%)	879 / 400	6,6	1,65%
Присутствует 2 поля или больше (77,3%)	875 / 400	6,08	1,52%

Результаты эксперимента



Характеристики алгоритма

- Работа с неполными данными
- Учет взаимной зависимости признаков
- Настройка на конкретную базу данных
- Расширенные авторитетные записи
- Перенос на задачу выявления дубликатов

Заключение

- Представлены алгоритм и технология автоматического связывания документов
- Планируется внедрение в распределенный электронный каталог медицинских библиотек НП «МедАрт»

Автоматическое связывание документов

Князева Анна

aknjazeva@ict.nsc.ru

Спасибо за внимание!