Сколковский Институт Науки и Технологий

Московский Физико-Технический Институт

(государственный университет)

Рекомендательная система подбора одежды, основанная на метрическом анализе рекламных описаний

Владимир Владимирович Жуйков

Научный руководитель:

д.ф.-м.н, н.с. ВЦ РАН

В. В. Стрижов

Научный руководитель:

Профессор Университета Карнеги Мелон, А. В. Гершман

Москва, 2015

Цель исследования

Цель исследования

Разработать рекомендательную систему подбора одежды.

Решаемая задача

На основе заданного пользователем текстового описания одежды построить систему ранжирования документов и вывести m ближайших результатов, при условии максимального расстояния между элементами выдачи. Определить важность признаков в смешанных шкалах.

Методы решения. Для построения ранжирующей модели – Heterogeneous Euclidean-Overlap Metric (HEOM) метрика, заданная в смешанных шкалах. В работе используется обучение ранжированию, используя Gradient descent and SQP алгоритмы оптимизации.

Описание данных

№	Признак	Шкала	Пример
1	Тип	nominal, ℂ	Платье
2	Подтип	nominal, ℂ	Платья-миди
3	Цена	linear, 𝕎	[0…+∞]
4	Цвет	nominal, ℂ	зеленый
5	Описание	text, 𝕋	Платье Lawiggi выполнено из атласного зеленого текстиля. Модель имеет приталенный крой, подчеркнутый пышной юбкой поверх с сеткой в тон.
6	Фотография	picture, ℙ	LA003EWDBP84.jpg
7	Бренд	nominal,ℂ	Lawiggi
8	Сезон	nominal, ℂ	Мульти
9	Коллекция	nominal, ℂ	Осень-зима
10	Страна производства	nominal, ℂ	Турция
11	Длина по спинке	linear, 𝕎	[0…200]
12	Длина рукава	linear, 𝕎	[0…200]
13	Детали одежды	nominal, ℂ	эффектный цветок на талии
14-30	17 материалов	linear, 𝕎	[0…100]
31	Артикул	nominal, ℂ	LA003EWDBP84
32-37	6 экспертных оценок	nominal, ℂ	{0,1} вечернее/повседневное, скромное/броское, взрослое/молодежное

1 текст, 20 линейных, 16 номинальных шкал 4435 документов

Обзор литературы

Теоретическая часть

- *Рекомендательные системы:* Francesco Ricci, et. al.: Recommender Systems. Handbook (2011).
- *HEOM:* Wilson D.R., et. al.: Improved heterogeneous distance functions (1997).
- *Методы оптимизации:* A. Nemirovski: Optimization numerical methods for nonlinear continuous optimization (1999).
- *Важность признаков:* Taher H. Haveliwala.: Topic-Sensitive PageRank (2002).

Обзор существующих рекомендательных систем подбора одежды

- *Сценарий реальной жизни:* Si Liu, et. al.: Hi, Magic Closet, Tell Me What to Wear! (2012).
- *Атрибуты стилей:* Wei Di2, et. al.: Style Finder: Fine-Grained Clothing Style Recognition and Retrieval (2013).
- *Интерактивное взаимодействие:* Lamche B., et. al.: Interactive Explanations in Mobile Shopping Recommender Systems (2014).

Постановка задачи

Исходная выборка представляет собой множество пар, описанных в смешанных шкалах:

$$\mathfrak{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}): i \in \mathcal{I}\}, i \in \mathcal{I}\{1, \dots, m\},$$

$\boldsymbol{x}_i$ - вектор i-го запроса пользователя, $\boldsymbol{x}_i \in \mathbb{X}$, $\mathbb{X} = \mathbb{L}_1 \times \dots \times \mathbb{L}_n$ множество возможных значений векторов признаков объектов,

$\boldsymbol{y}_i$ - вектор результатов запроса $\boldsymbol{x}_i$; $\boldsymbol{y}_i = [\boldsymbol{y}_i^1, \dots, \boldsymbol{y}_i^k], \boldsymbol{y}_i \in \mathbb{X}^{\boldsymbol{k}}$.

Объекты описаны в смешанных шкалах с заданной метрикой d:

$$d: \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+,$$

$$d(\boldsymbol{x}_i, \boldsymbol{y}_i^j) = \sqrt{\sum_{q=1}^{n} \alpha_q \cdot r^2(x_{iq}, y_{iq}^j)},$$

где n - количество признаков, j $\in \mathcal{J}\{1, \dots, k\}$, α_q - весовой коэффициент признака q, $r(x_{iq}, y_{iq}^j)$ – расстояние между векторами $\boldsymbol{x}_i$, $\boldsymbol{y}_i^j$ признака q.

Постановка задачи

Требуется найти коэффициенты $(a_1, \dots, a_n)$, удовлетворяющие следующим условиям:

$$\boldsymbol{\alpha} = (a_1, \dots, a_n) = \begin{cases} \operatorname{argmin} \sum_{i=1}^{m} \sum_{\boldsymbol{y}_i \in \mathbb{Y}_r} d(\boldsymbol{x}_i, \boldsymbol{y}_i), \\ \operatorname{argmax} \sum_{i=1}^{m} \sum_{\boldsymbol{y}_i \in \mathbb{Y}_{nr}} d(\boldsymbol{x}_i, \boldsymbol{y}_i), \\ \operatorname{argmax} \sum_{i=1}^{m} \sum_{\boldsymbol{y}_i', \boldsymbol{y}_i'' \in \mathbb{Y}_r} d(\boldsymbol{y}_i', \boldsymbol{y}_i''), \end{cases}$$

где m - количество запросов, $\mathbb{Y}_r$ - множество релевантных ответов, $\mathbb{Y}_{nr}$ - множество нерелевантных ответов; y, y', y'' - результаты запроса.

Утверждение 1. Так как d - метрика, поэтому $d \geq 0$. Так как $r^2(x_{iq}, y_{iq}^j) \geq 0$, поэтому $\alpha_q \geq 0$. $\sum_{q=1}^{n} a_q = 1$

МФТИ

Постановка задачи

$$A = \frac{\sum_{i=1}^{m} \frac{\sum_{\boldsymbol{y}_i \in \mathbb{Y}_r} d_i(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\sqrt{n} \cdot |\mathbb{Y}_r|}}{m}, B = \frac{\sum_{i=1}^{m} \frac{\sum_{\boldsymbol{y}_i \in \mathbb{Y}_{nr}} d_i(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\sqrt{n} \cdot |\mathbb{Y}_{nr}|}}{m}, C = \frac{\sum_{i=1}^{m} \frac{2 \cdot \sum_{\boldsymbol{y}'_i, \boldsymbol{y}''_i \in \mathbb{Y}_r} d_i(\boldsymbol{y}'_i, \boldsymbol{y}''_i)}{\sqrt{n} \cdot |\mathbb{Y}_r \cdot (|\mathbb{Y}_r| - 1)|}}{m}.$$

Значения $A, B, C \in [0,1]$, поэтому мы можем переписать:

$$\boldsymbol{\alpha} = (a_1, \dots, a_n) = \begin{cases} \operatorname{argmin} A(\boldsymbol{\alpha}), \\ \operatorname{argmax} B(\boldsymbol{\alpha}), \\ \operatorname{argmax} C(\boldsymbol{\alpha}). \end{cases}$$

Следовательно, необходимо решить оптимизационную задачу: поиск максимума функции $f(\boldsymbol{\alpha})$:

$$f(\boldsymbol{\alpha}) = B(\boldsymbol{\alpha}) \cdot \big(1 - A(\boldsymbol{\alpha})\big) + \lambda \cdot C(\boldsymbol{\alpha}) \to \max,$$

где $\lambda = \text{const}$

HEOM метрика для смешанных шкал

Шкала – алгебраическая структура с заданным набором операций и отношений, удовлетворяющая фиксированным набором аксиом.

Номинальная шкала – шкала, с введенным на ней отношением, удовлетворяющим аксиомам тождества.

Линейная шкала – порядковая шкала с заданными на ней операциями сложения и вычитания.

HEOM:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{q=1}^{n} \alpha_q \cdot r^2(x_{iq}, x_{jq})}$$

Расстояния:

$$r(x_{iq}, x_{jq}) = \begin{cases} \text{diff}\,(x_{iq}, x_{jq}), \text{если } \mathbb{L}_q - \text{линейная шкала}, \\ \text{overlap}(x_{iq}, x_{jq}), \text{если } \mathbb{L}_q - \text{номинальная шкала}, \\ \text{sim}\,(x_{iq}, x_{jq}), \quad \text{если } \mathbb{L}_q - \text{текст}, \end{cases}$$

$$\text{diff}(x_{iq}, x_{jq}) = \frac{|x_{iq} - x_{jq}|}{max_{\mathbb{L}_q} - min_{\mathbb{L}_q}},$$

$$\text{sim}(x_{iq}, x_{jq}) = \arccos \frac{< x_{iq}, x_{jq} >}{||x_{iq}|| \cdot ||x_{jq}||}.$$

$$\text{overlap}(x_{iq}, x_{jq}) = \begin{cases} 1, \text{если } x_{iq} \neq x_{jq}, \\ 0, \quad \text{иначе}, \end{cases}$$

МФТИ

Вычислительный эксперимент

Цель вычислительного эксперимента

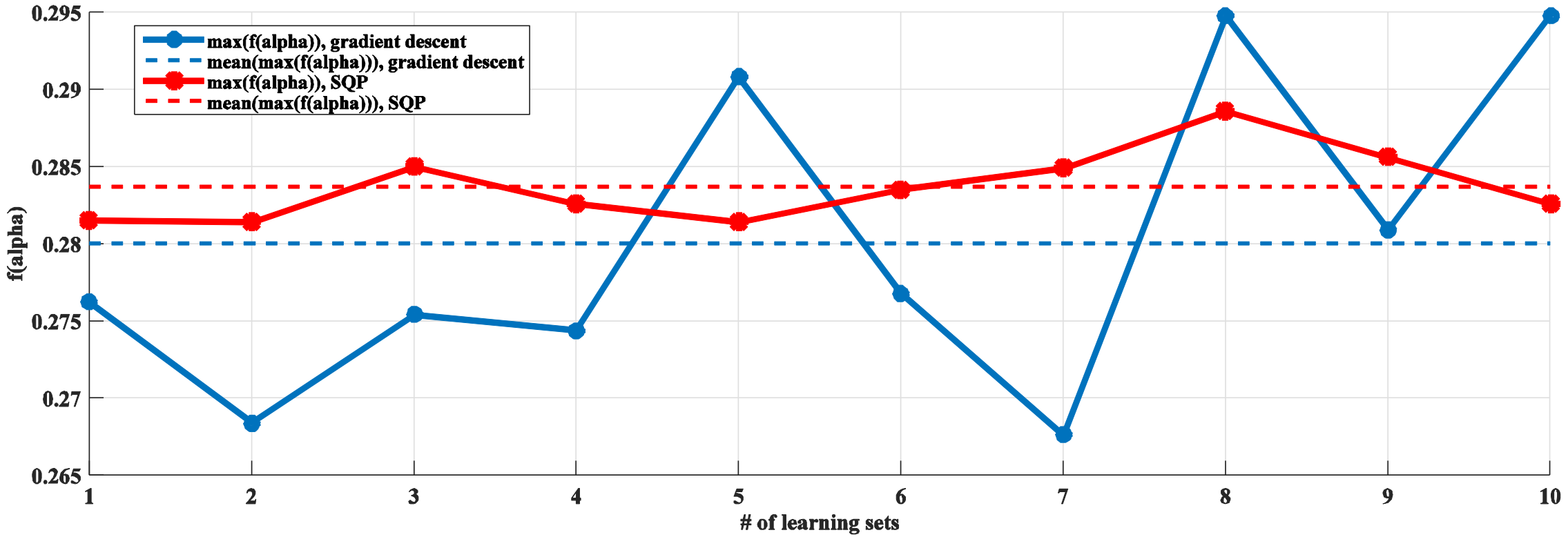
- Решить оптимизационную задачу.
- Проверить адекватность предложенной метрики при решении задачи обучения ранжированию.

Описание данных

- *Число записей:* 4435 документов.
- *Число признаков:* 35 (1 текст, 20 линейных, 14 номинальных).
- 200 запросов с 20 отранжированными релевантными и нерелевантными результатами запроса (экспертные оценки).
- 200x20 = 4000 – мощность выборки.
- *Набор на обучение:* 100 запросов (10 раз).
- *Набор на тестирование и контроль:* 100 запросов (10 раз).

Стандартное отклонение функции ошибки

Максимальные значения функции ошибки f($\boldsymbol{\alpha}$), используя GD (градиентный спуск) и SQP (последовательное квадратичное программирование).



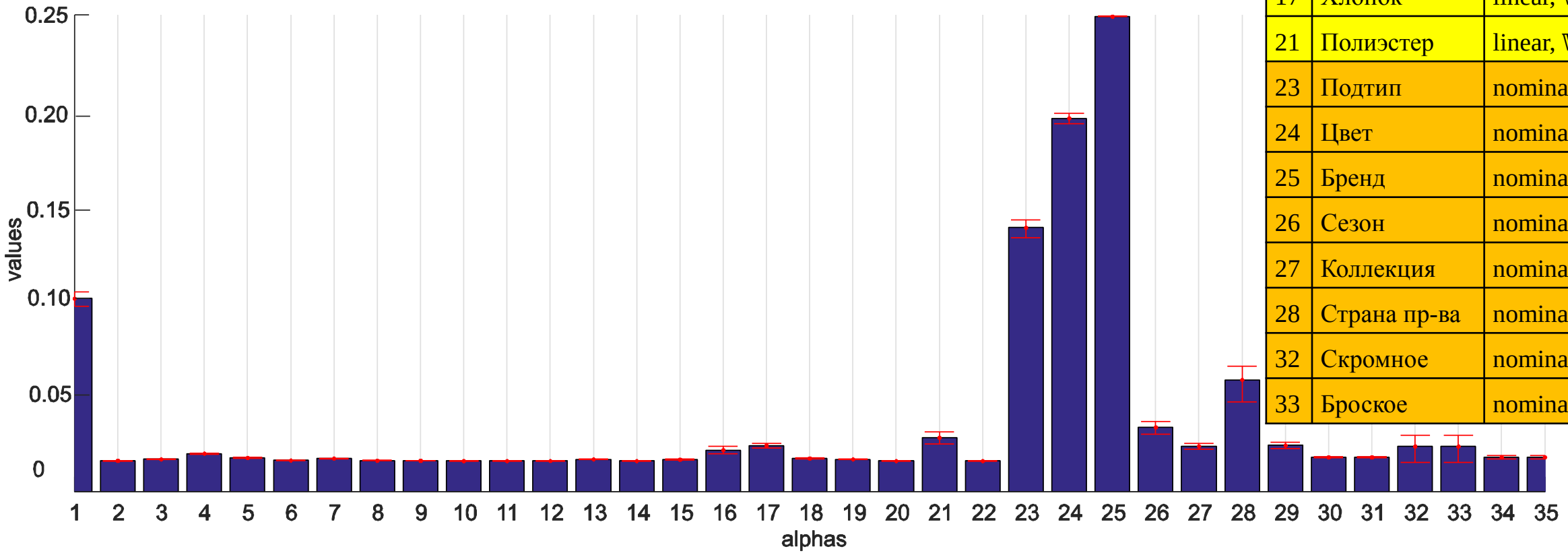
$$\text{GD: } \max f(\boldsymbol{\alpha}) = \overline{\max f(\boldsymbol{\alpha})} \pm \delta = \ 0.2800 \pm 0.0101$$

$$\text{SQP: } \max f(\boldsymbol{\alpha}) = \overline{\max f(\boldsymbol{\alpha})} \pm \delta = \ 0.2837 \pm 0.0023$$

МФТИ

Важность признаков

Важность признаков, описанных в смешанных шкалах (SQP–стандартное отклонение)



№	Признак	Шкала
1	Описание	text, 𝕋
4	Длина рукава	linear, 𝕎
16	Акрил	linear, 𝕎
17	Хлопок	linear, 𝕎
21	Полиэстер	linear, 𝕎
23	Подтип	nominal, ℂ
24	Цвет	nominal, ℂ
25	Бренд	nominal,ℂ
26	Сезон	nominal, ℂ
27	Коллекция	nominal, ℂ
28	Страна пр-ва	nominal, ℂ
32	Скромное	nominal, ℂ
33	Броское	nominal, ℂ

Критерии качества

- **Precision, точность, P@20.** $(\mathrm{P@20})_m = \frac{|\mathbb{Y}_r^m|}{20}$, $m = 1 \ldots 100$,
- **MAP – средняя точность по позициям релевантных документов по всем запросам**

$$\mathrm{MAP} = \frac{1}{m} \cdot \sum_{j=1}^{m} \frac{1}{|\mathbb{Y}_r^j|} \cdot \sum_{i=1}^{|\mathbb{Y}_r^j|} P(\mathrm{doc}_i)$$

m – количество запросов, $|\mathbb{Y}_r^j|$ – количество релевантных ответов запросу j, $P(\mathrm{doc}_i)$ – точность i-го релевантного документа

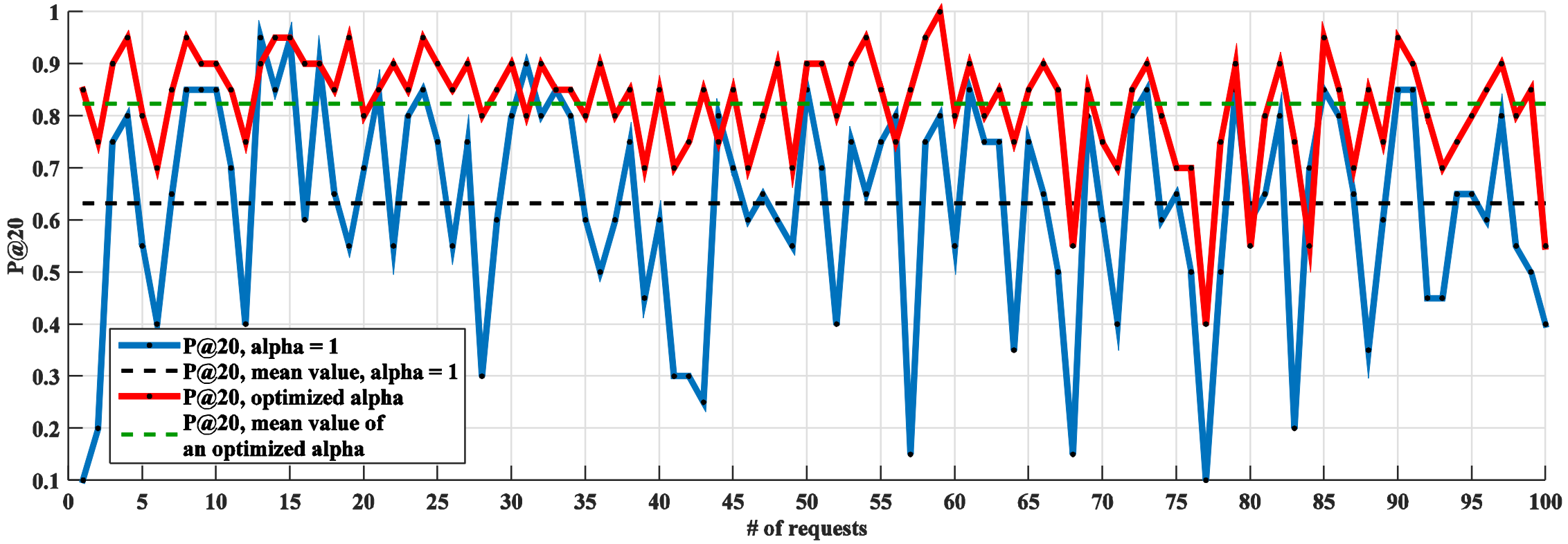
- **nDCG – нормированная дисконтированная сумма выигрышей**

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)},$$

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG_p}},$$

$IDCG_p$ – DCG_p при идеальном ранжировании

P@20



До оптимизации:

min = 0.1000 mean = 0.6315

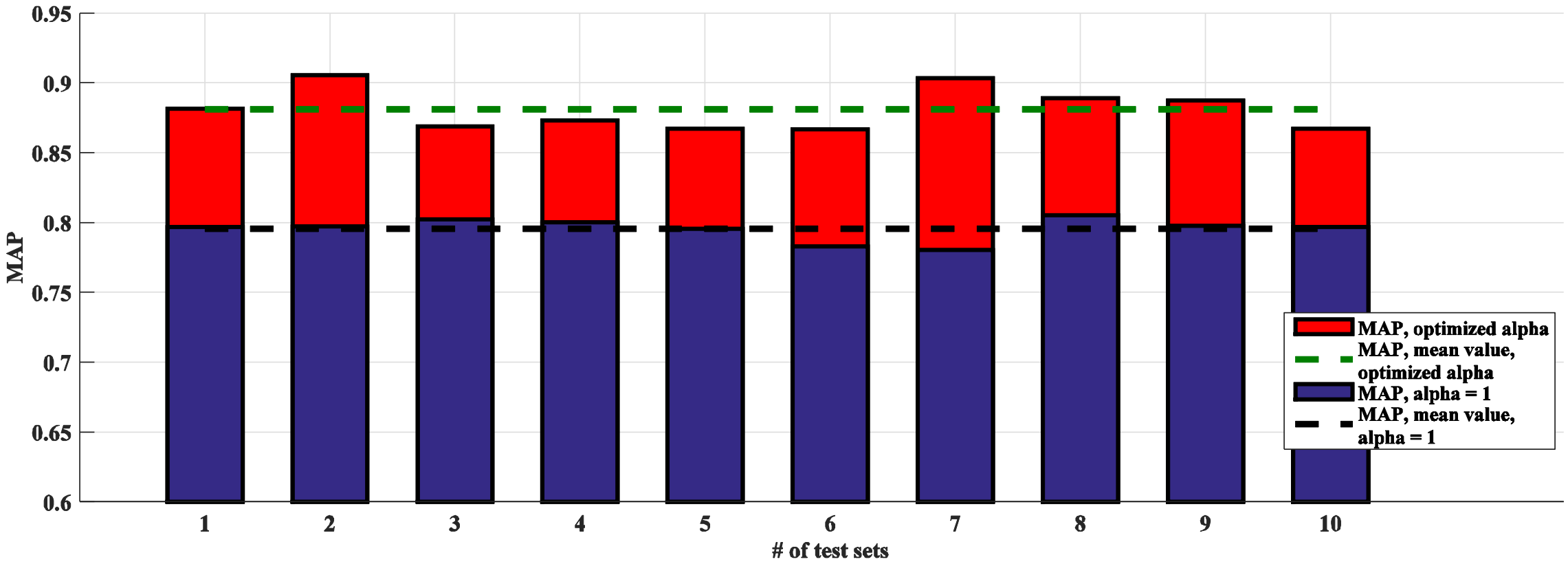
max = 0.9500

После оптимизации:

min = 0.4000 mean = 0.8230

max = 1.0000

Mean average precision



До оптимизации:

min = 0.7805 mean = 0.7956

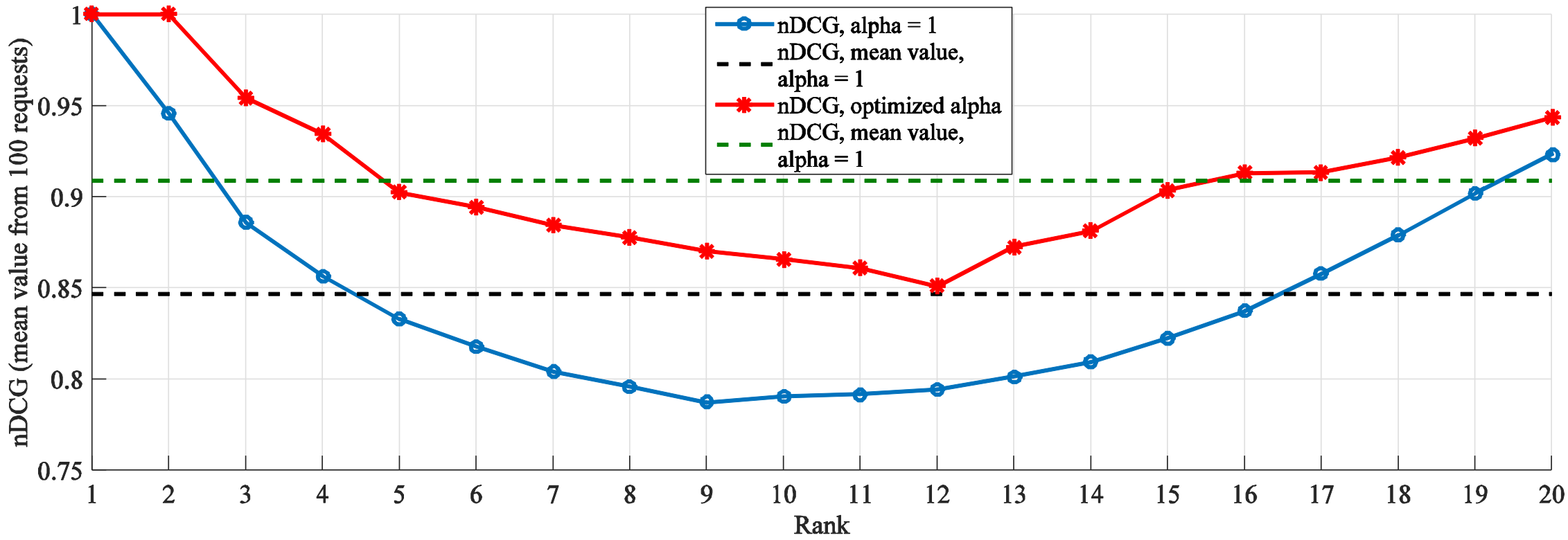
max = 0.8055

После оптимизации:

min = 0.8668 mean = 0.8811

max = 0.9055

Normalized Discounted Cumulative Gain



До оптимизации:

min = 0.7872 mean = 0.8467

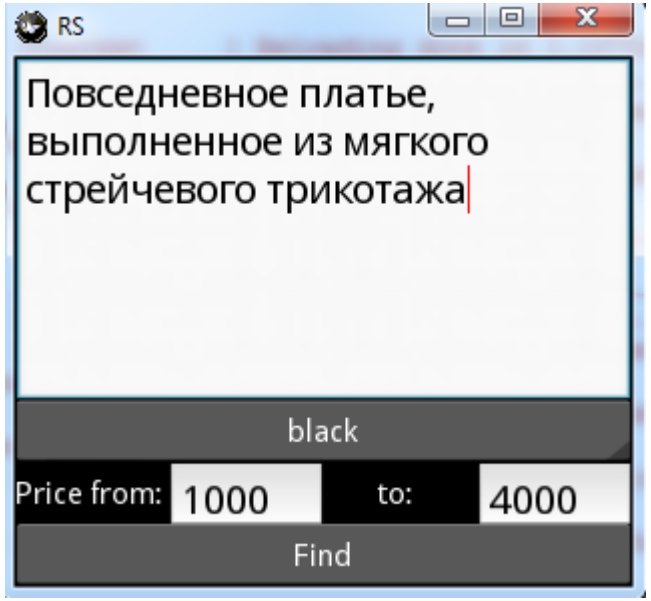
max = 1.0000

После оптимизации:

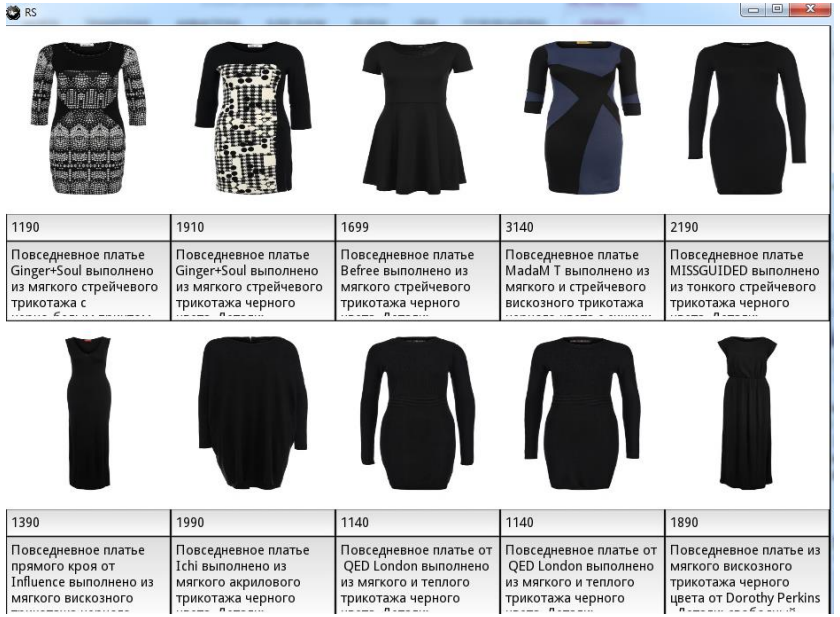
min = 0.8508 mean = 0.9088

max = 1.0000

Прототип

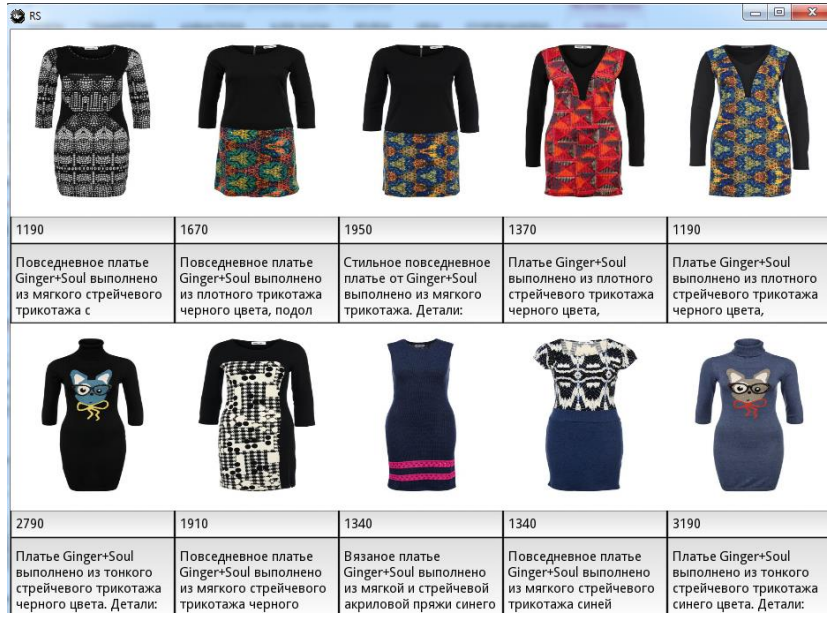


Window 1: Запрос пользователя



Window 2: *k* результатов запроса

Пользователь выбирает 1 платье



Window 3: *m* результатов запроса Windows 2

Demo: https://youtu.be/FSFhTde6iO8

МФТИ

Заключение

- Разработана рекомендательная система подбора одежды, основанная на метрическом анализе описаний одежды
- Поставлена и решена минимаксная задача ранжирования объектов в смешанных шкалах.
- Разработан алгоритм метрического ранжирования объектов по запросу.
- Определена важность признаков (наиболее важные признаки для пользователей: «Описание», «Акрил», «Хлопок», «Подтип», «Цвет» и «Бренд»).
- Разработан прототип рекомендательной системы подбора одежды (техническая система).

Appendix – Recommendation system

Recommendation system is a software tool and techniques that provide useful suggestions for users.

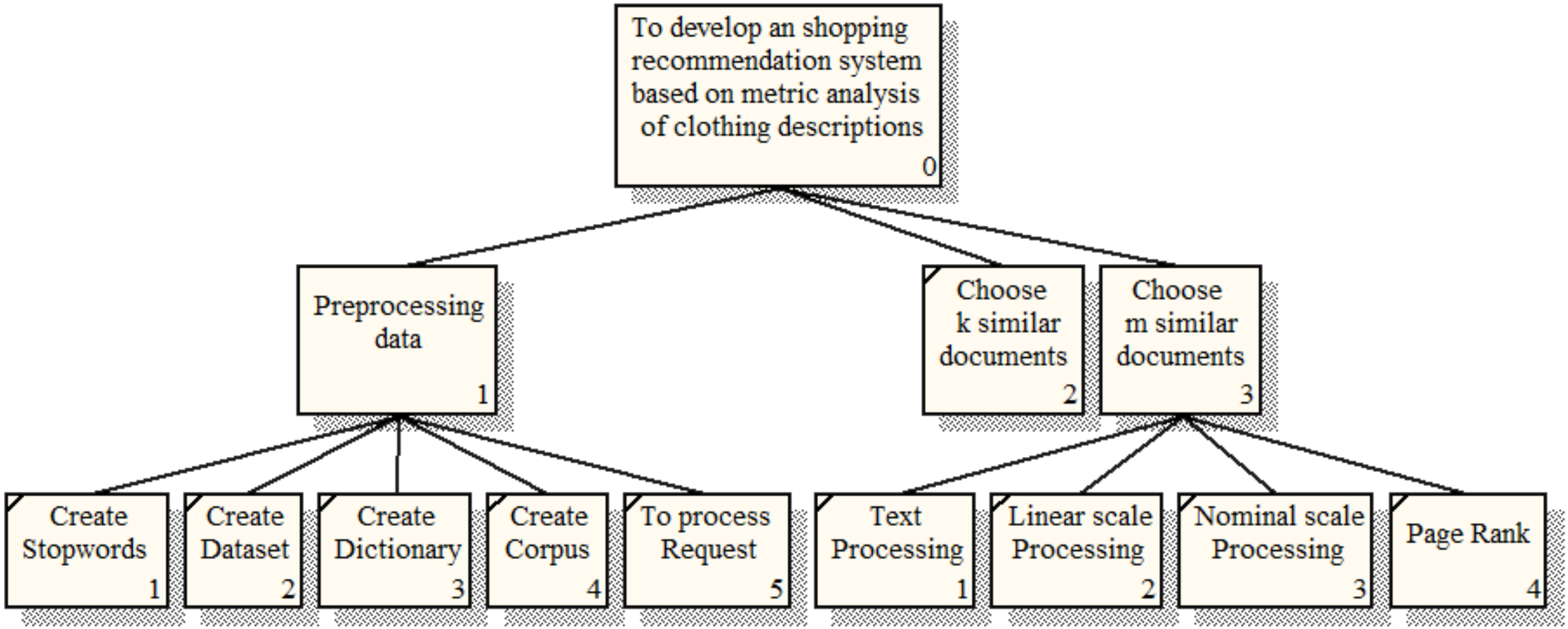
Types: collaborative filtering, content-based filtering, hybrid techniques, mobile

Clothing recommendation system based on: reasonable computing, concrete attributes, web mining, the “wisdom of crowds”, multimedia mining, active learning strategy, photos from magazines, scenarios

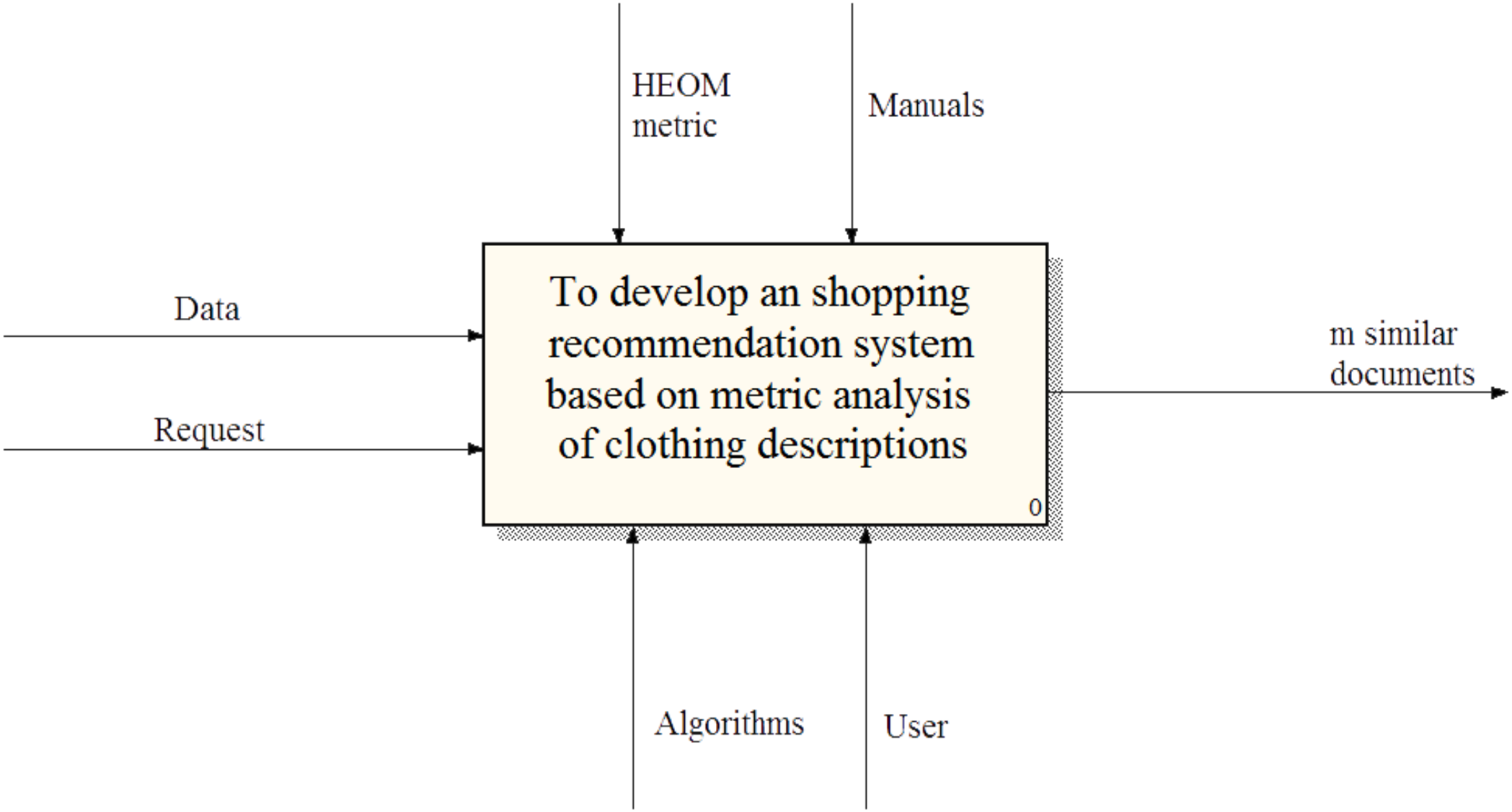
Technology requirement

- Euclidian distance function
- Heterogeneous Euclidean-Overlap Metric (HEOM)
- Frequency–Inverse Document Frequency (TF.IDF)
- Cosine similarity measure
- Gradient descent
- Sequential Quadratic Programming (SQP)

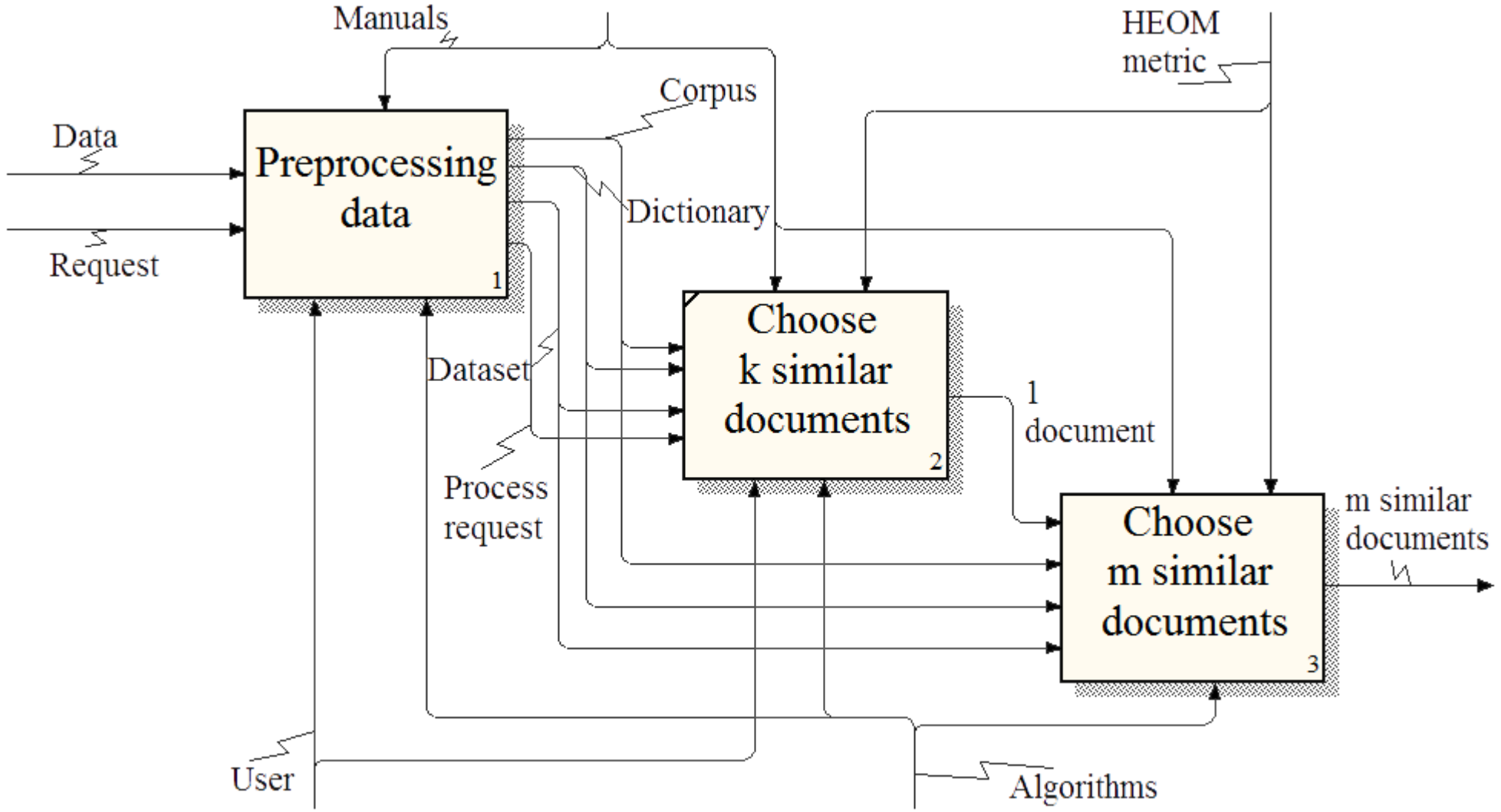
Appendix – Node tree diagram



Appendix

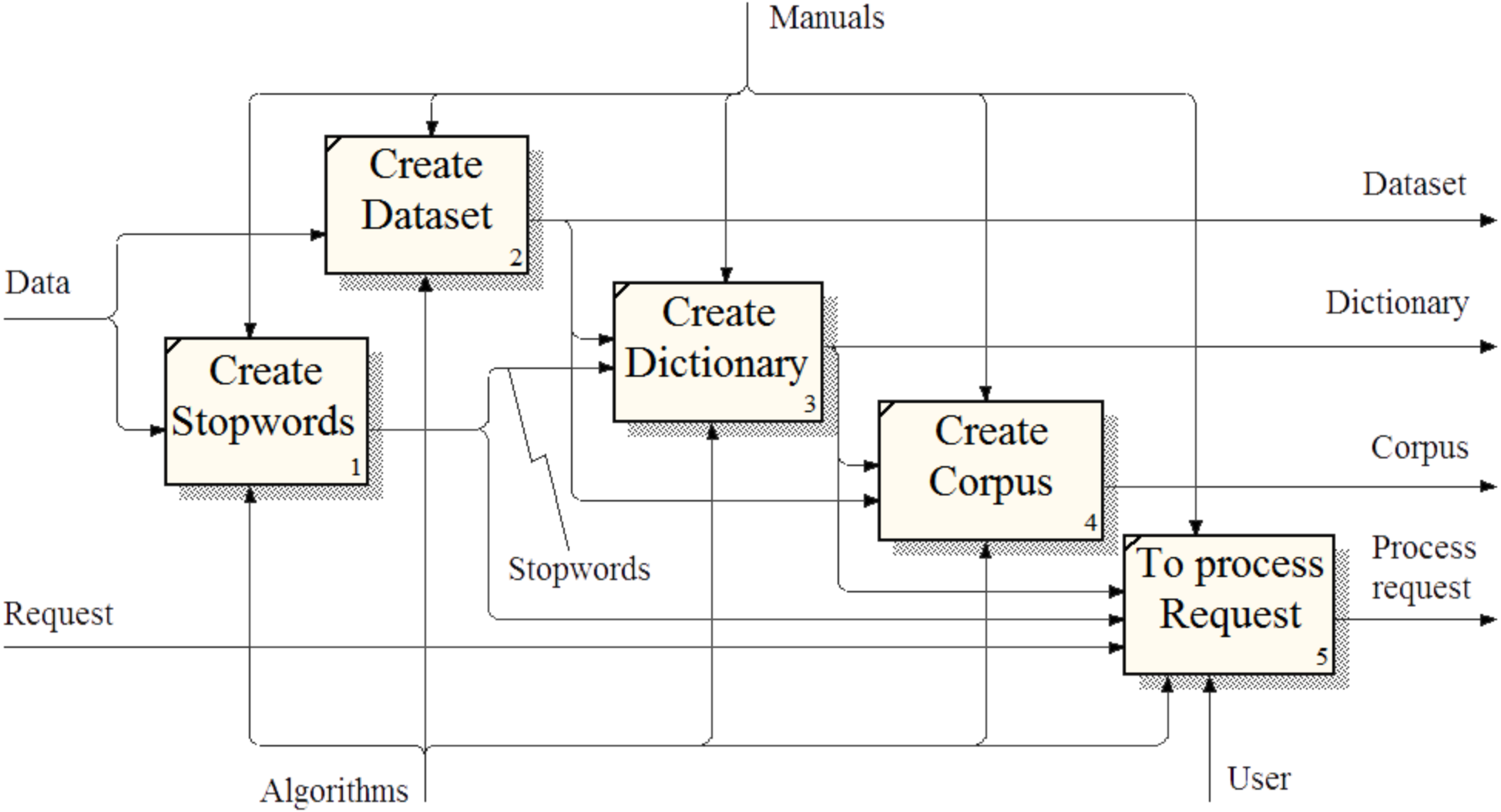

HEOM metric
Manuals
Data
Request
To develop an shopping recommendation system based on metric analysis of clothing descriptions
0
m similar documents
Algorithms
User

Appendix – IDEF0 – the first level of the decomposition



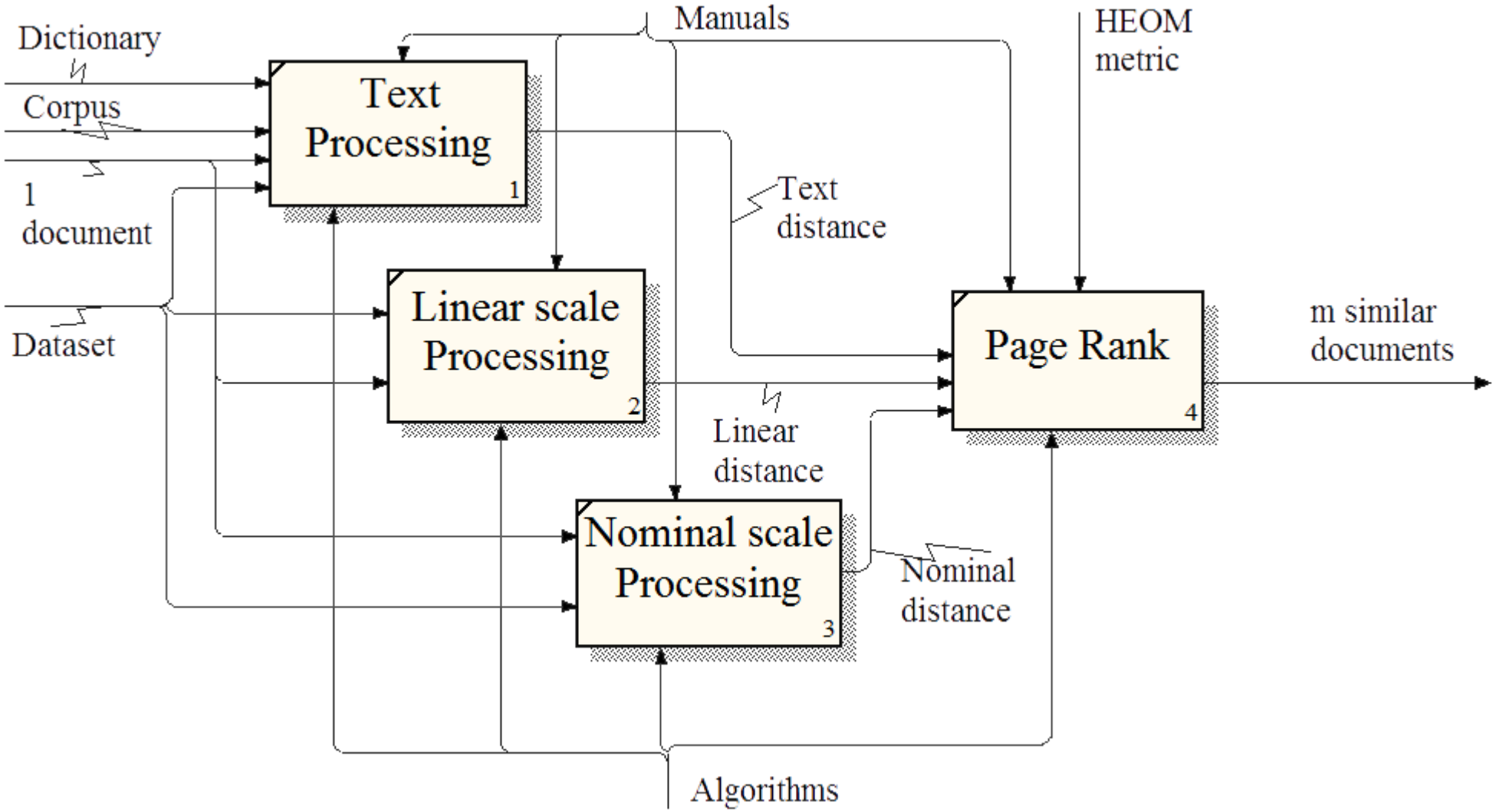
The first level of the decomposition “To develop a shopping recommendation system based on metric analysis of clothing descriptions”

Appendix – IDEF0 – decomposition “Preprocessing data”



The second level of the decomposition “Preprocessing data”

Appendix – decomposition “Choose *m* similar documents”



The second level of the decomposition “Choose *m* similar documents”

Appendix

Function words	Number
Interjections	302
Particles	151
Prepositions	189
Pronouns	38
Question words	16
Unions	141

Appendix

Consider a nonlinear programming problem of the form:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & b(x) \geq 0 \\ & c(x) = 0. \end{aligned}$$

The Lagrangian for this problem is;

$$\mathcal{L}(x, \lambda, \sigma) = f(x) - \lambda^T b(x) - \sigma^T c(x),$$

where λ and σ are Lagrange multipliers. At an iterate x_k, a basic sequential quadratic programming algorithm defines an appropriate search direction d_k as a solution to the quadratic programming subproblem

$$\begin{aligned} \min_{d} \quad & f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T \nabla^2_{xx} \mathcal{L}(x_k, \lambda_k, \sigma_k) d \\ \text{s.t.} \quad & b(x_k) + \nabla b(x_k)^T d \geq 0 \\ & c(x_k) + \nabla c(x_k)^T d = 0. \end{aligned}$$

Note that the term $f(x_k)$ in the expression above may be left out for the minimization problem, since it is constant.

Appendix

input: $f(x)\colon \mathbb{R}^n \to \mathbb{R}$

output: optimum x

1) repeat: $x^{[k+1]} = x^{[k]} - \lambda^{[k]} \cdot \nabla f\left(x^{[k]}\right)$, where $\lambda^{[k]}$:
   - const ($f(x)$ – differentiable, bounded above or strongly convex with const A),
   - decreases with fractional step (when *const* method does not work),
   - $\lambda^{[k]} = \underset{\lambda}{\operatorname{argmin}} f\left(x^{[k]} - \lambda \cdot \nabla f\left(x^{[k]}\right)\right)$ (steepest descent method).
2) if the stopping criterion holds, then output = $x^{[k+1]}$

Stopping criterion:

1) $\left\|x^{[k+1]} - x^{[k]}\right\| \leq \epsilon$,
2) $\left\|f(x^{[k+1]}) - f(x^{[k]})\right\| \leq \epsilon$, ϵ – *const.*