طراحی و پیاده سازی الگوریتم تطابق اسامی در زبان فارسی به منظور تشخیص ذینفع واحد

لیلا مومنی نسب، momeninasab.leila@gmail.com
دکتر نیما امیرشکاری، nima.itpro@gmail.com
استاد جلال ملکی، jalal.maleki@liu.se
پرفسور لارش اهرنبرگ، lars.ahrenberg@liu.se

سومین همایش سالانه بانکداری الکترونیک و نظام های پرداخت

conf.mbri.ac.ir/ebps3

کاربرد
مروری بر الگوریتم های موجود
یک الگوریتم تطابق اسامی برای زبان فارسی
پی دا
ارزیابی
نتیجه
آینده کاری

کاربرد

- در زمینهٔ بانکداری و مالی برای تشخیص تقلب
- مدیریت ارتباط با مشتری
- ضد پولشویی
- رتبه بندی اعتباری
- تشخیص ذینفع واحد

مروری بر الگوریتم های موجود

- تغییرات اسامی
- الگوریتم های تطابق اسامی
- زبان فارسی

مروری بر الگوریتم های موجود
تغییرات اسامی

در اسامی یکسان اتفاق می افتد

- تغییرات نوشتاری
  - ✓ خطاهای املائی
  - ✓ نوشتار جایگزین
  - ✓ نویسه گردانی
  - ✓ حروف بی صدا
- تغییرات فیلد
- اسامی هم ارز

مروری بر الگوریتم های موجود
تغییرات اسامی

- مختصر نویسی
- قطعه شدن
- ترجمه
- اضافه یا حذف اجزاء
- استفاده از علائم

مروری بر الگوریتم های موجود
الگوریتم های تطابق اسامی

آیا این دو اسم به یک شخص واحد تعلق دارد؟

Mohamed Amin و Mohamedamin

- الگوریتم های تطابق صوتی
  - Soundex
- الگوریتم های تفاوت رشته ای
  - ***Levenshtein***
- الگوریتم های بر مبنای توکن
  - Q-gram

مروری بر الگوریتم های موجود
زبان فارسی

- الفبای زبان فارسی
  - ۳۳ حرف
  - سیستم نوشتاری Perso-Arbic
  - از چپ به راست
  - اتصال حروف در نوشتار

مروری بر الگوریتم های موجود
یک الگوریتم تطابق اسامی برای زبان فارسی

- Levenshtein ➡ Arabic Edit Distance Algorithm (AEDA)

  ➡ *PEDA*

پی دا

این دو اسم فارسی چه اندازه با هم شباهت دارند؟

موراذ و مراد

پی دا

- سطوح شباهت در زبان فارسی
  - ✓ شباهت فرمی
  - ✓ شباهت صوتی
  - ✓ شباهت کی بوردی
- هستهٔ اصلی کد پی دا

پی دا
سطوح شباهت در زبان فارسی
شباهت فرمی

Form Similarity in Persian Alphabet(between origin letter forms)		
No.	Similar Groups	Similarity Index
1.	(ی - ي) (ه - ۀ - ة - ہ)(أ - إ - ا) (أ - إ - ا) (ا - أ) (ا - إ) (آ - إ) (ک - ك)	۱
2.	(ؤ - و) (ت - ٹ) (ج - چ) (ح - خ) (د - ذ) (ر - ز) (ص - ض) (ط - ظ) (ع - غ) (ک - گ) (ی - ئ)	۰.۸
3.	(ب - ی) (ب - ي) (پ - ی) (پ - ي) (ت - ن) (ف - ق)	0.54
4.	(ب - پ) (چ - ڇ) (ز - ژ)	۰.٦
5.	(چ - ح) (س - ش) (ر - ژ) (ٹ - ن)	۰.٤
6.	(غ - ف)	0.27
7.	(ك - ل) (ب - ت) (ب - ٹ) (پ - ت) (پ - ٹ) (ج - خ) (چ - خ)	۰.۲
8.	(ب - ن) (پ - ن) (ٹ - ی) (ٹ - ي) (ن - ی) (ن - ي) (ت - ی) (ت - ي) (ئ - ت) (ئ - ٹ) (ئ - ن)	0.14
9.	(م - ه)	0.07
10.	Any other pair of Persian letters	0

پی دا
سطوح شباهت در زبان فارسی
شباهت صوتی

Phonetic Similarity in Persian Alphabet(between origin letter forms)		
No.	Similar Groups	Similarity Index
	(ک - ك) (ی - ى) (أ - ا) (أ - ع) (أ - ى) (ا - ا) (ؤ - و) (ت - ط) (ث - س - ص) (ز -ذ-ض-ظ) (ح - ه) (ح - ة) (ع - ا) (غ - ق) (ؤ - ع) (ة - ه)	۱
	(ئ - ی)	۰.۸
		۰.٦
	(ب - پ) (ت - د) (ث،س،ص - ز،ذ،ض،ظ) (ج - چ) (ژ - ش) (ف - و) (ک - گ) (ك - گ) (م - ن)	۰.٤
		۰.۲
	(ة - ت)	۰.۱
	Any other combination of Persian letters	0

پی دا
سطوح شباهت در زبان فارسی
شباهت کی بوردی

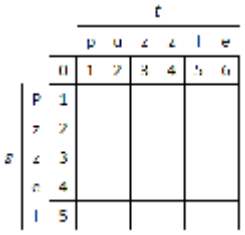
$$Sim_{kb}(a,b) = 1 - \frac{\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}}{\psi}$$

پی دا
هستهٔ اصلی کد پی دا

- Levenshtein
  - ✓ تبدیل اسم مبدا به اسم مقصد با کمترین هزینه
  - ✓ حداقل تعداد عملیات برای تبدیل
    - جایگزینی
    - حذف
    - اضافه

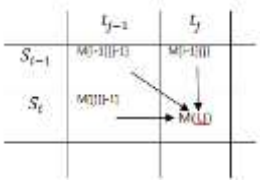
پی دا
هستهٔ اصلی کد پی دا

- Levenshtein
  - ✓ یک ماتریکس می سازد
  - ✓ سطر اول و ستون اول ماتریکس را پر می کند



پی دا
هستهٔ اصلی کد پی دا

- Levenshtein
  - ✓ سلول های ماتریکس را با توجه به مقادیر همسایه ها پر می کند



		p	u	z	z	l	e
	0	1	2	3	4	5	6
p	1	0	1	2	3	4	5
z	2	1	1	1	2	3	4
z	3	2	2	1	1	2	3
e	4	3	3	2	2	2	2
l	5	4	4	3	3	2	3

$$M(i,j) = min \begin{cases} M(i-1,j-1) + substitution\ cost \\ M(i-1,j) + delete\ cost \\ M(i,j-1) + insert\ cost \end{cases}$$

پی دا
هستهٔ اصلی کد پی دا

- هزینهٔ عملیات حذف و اضافه

محمد عرفان
انشا
محمد
اسماعیل

محمد عرفان
انشا
محمد
اسماعیل

$$f_{del}(a_i, a_{i-1}) = \begin{cases} 0 & if\ a_i = blank \\ 0 & if\ a_i \neq a_{i-1} \wedge a_i = * \wedge a_{i-1} \in \xi = \{alif, ya, waw\} \\ \theta & if\ a_i = a_{i-1} \\ \mu & if\ a_i \in \xi = \{alif, ya, waw\} \\ 1 & Otherwise \end{cases}$$

- هزینهٔ عمل جایگزینی

$$f_{sw}(a,b) = \begin{cases} \dfrac{\alpha(a,b).\omega + \beta(a,b).\lambda + \gamma(a,b).\sigma}{\omega + \lambda + \sigma} & if\ a \neq b \\ 0 & Otherwise \end{cases}$$

پی دا
هستهٔ اصلی کد پی دا

- مثال

موراذ و مراد

		م	و	ر	ا	د
	0	1	2	3	4	5
م	1	0	0.15	1.15	1.3	2.3
ر	2	1	0.77778	0.15	0.3	1.3
ا	3	1.15	0.92778	0.3	0.15	1.00595
د	4	2.15	1.87222	1.3	0.99444	0.57778

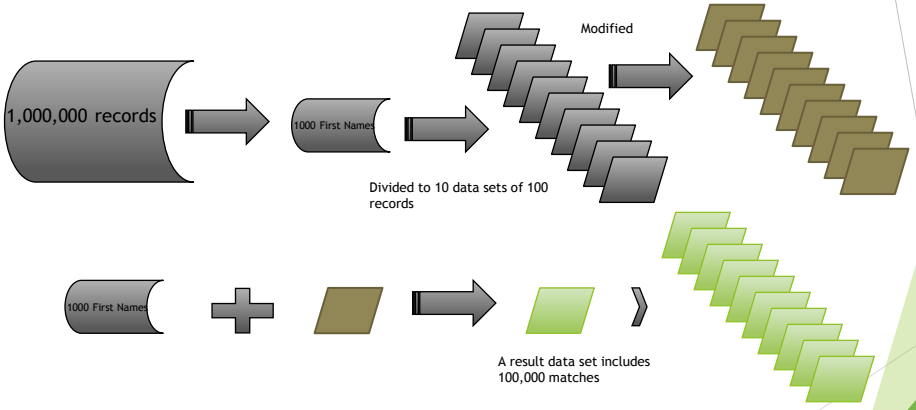
$$Sim(\text{موراذ, مراد}) = 100 - \frac{0.57778 * 100}{\max(4,5)} = 88\%$$

ارزیابی

- انجام سه سری از آزمایشات
- نتایج تطابق

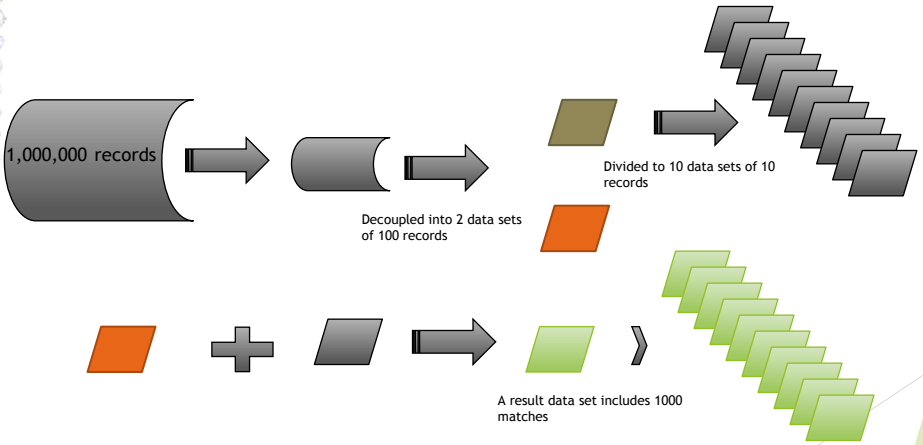
ارزیابی
انجام سه سری از آزمایشات

- اولین سری



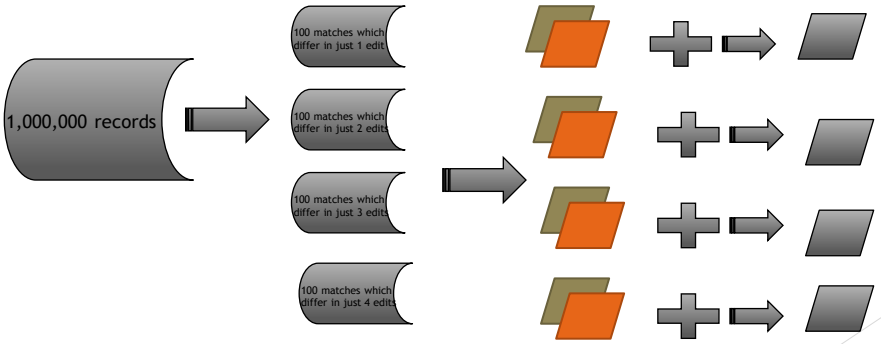
ارزیابی
انجام سه سری از آزمایشات

- سری دوم



ارزیابی
انجام سه سری از آزمایشات

- پی دا برای تعداد مختلفی از عملیات چگونه عمل می کند؟



ارزیابی
نتایج تطابق

- اولین سری

PEDA	Precision	Recall	f-measure
DS01	0.8	0.9	0.85
DS02	0.77	0.95	0.85
DS03	0.89	0.95	0.92
DS04	0.81	0.97	0.88
DS05	0.73	0.96	0.83
DS06	0.79	0.95	0.86
DS07	0.79	0.99	0.88
DS08	0.79	0.96	0.87
DS09	0.75	0.95	0.84
DS10	0.7	0.95	0.81
f-measure mean			0.86

ارزیابی
نتایج تطابق

- اولین سری: مقایسه با لونشتین

Levenshtein	DS01	DS02	DS03	DS04	DS05	DS06	DS07	DS08	DS09	DS10	
True positives	95%	100%	94%	99%	99%	96%	97%	96%	97%	98%	
False positives	5%	0%	6%	1%	1%	4%	3%	4%	3%	2%	
True negatives	59%	65%	65%	74%	67%	65%	69%	68%	62%	67%	
False negatives	41%	35%	35%	26%	33%	35%	31%	32%	38%	33%	

PEDA	DS01	DS02	DS03	DS04	DS05	DS06	DS07	DS08	DS09	DS10	
True positives	86%	78%	90%	81%	73%	79%	79%	21%	76%	70%	
False positives	14%	22%	10%	19%	27%	21%	21%	79%	24%	30%	
True negatives	92%	91%	91%	96%	93%	92%	98%	94%	96%	93%	
False negatives	8%	9%	9%	4%	7%	8%	2%	6%	4%	7%	

ارزیابی
نتایج تطابق

- دومین سری

	Precision	Recall	f-measure
DS01	1	1	1
DS02	0.64	0.78	0.7
DS03	0.83	0.55	0.66
DS04	0.75	0.43	0.55
DS05	0.71	0.83	0.76
DS06	0.86	1	0.92
DS07	1	1	1
DS08	1	0.75	0.86
DS09	1	0.71	0.83
DS10	0.71	0.83	0.76
f-measure mean			0.80

ارزیابی
نتایج تطابق

- سومین سری

	1 edit	2 edits	3 edits	4 + edits
True positives	99%	81%	69%	42%
False positives	1%	19%	31%	58%

ارزیابی

- انواع تغییرات اسامی را در داده های تست آورده ایم
- نتایج را برای انواع تغییرات اسامی بررسی نمودیم

s	*t*	Similarity
ح ح پری	پري	45%
زري	زهرا	68%
سید احمد	احمد	70%
نسیب آقا	نسیب	69%
اقا شهزاد	شهزاد	64%
مجحوبه	مجحوبه	76%
زینب نه نه خانم	زینب ن ن خا	68%
فاطمه	فاطي	64%

نتیجه

- پی دا برای تغییرات نوشتاری و قطعه شدن خوب عمل می کند
- توصیه می شود در کنار الگوریتم های دیگر استفاده شود

آینده کاری

- توسعهٔ قوانین شباهت
- هوشمند سازی پی دا
- ترکیب با پایگاه داده های موارد خاص (مانند اسامی اشخاص، شرکت ها، نام نامه و ...)

با تشکر

سومین همایش سالانه بانکداری الکترونیک و نظام های پرداخت

۱۶ و ۱۷ دی ماه ۱۳۹۲ - مرکز همایش های برج میلاد

conf.mbri.ac.ir/ebps3