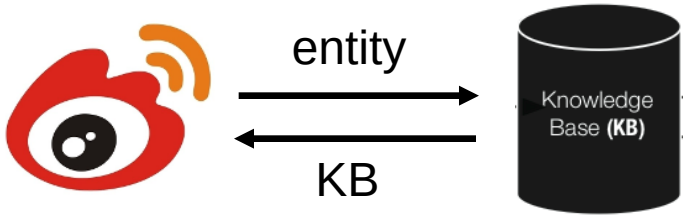
NLPCC 2014
The 3rd CCF Conference on Natural Language Processing & Chinese Computing
December 5-9,2014
Shenzhen

Entity Linking and Disambiguation in Chinese Micro-blogs



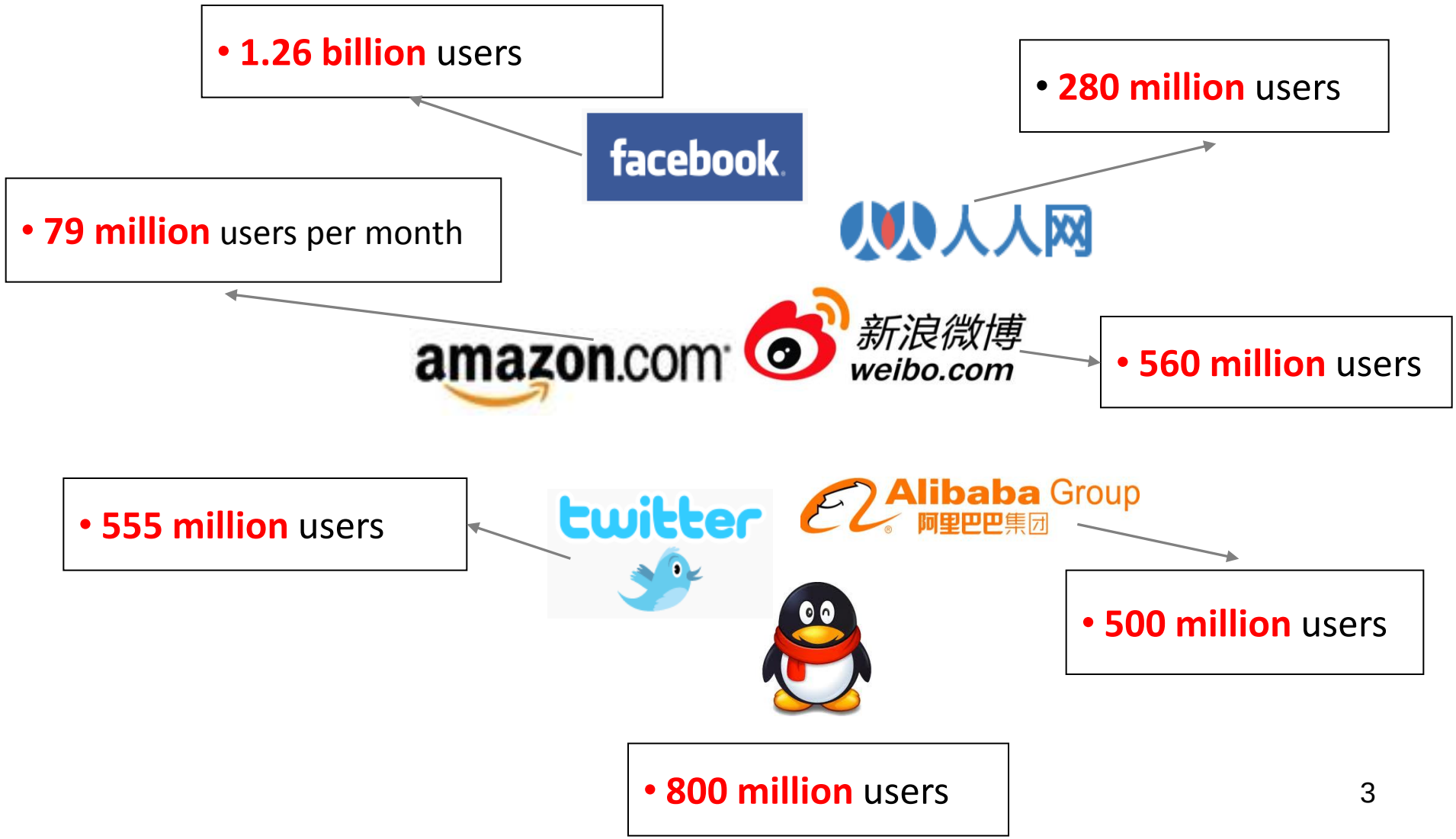
JinLan Fu, You He,
LinFeng Tian, Ning Fan, Li Li

Southwest University

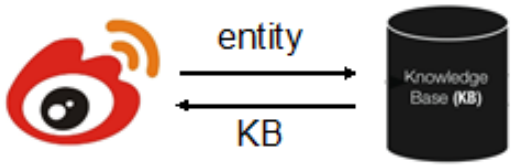
Motivation


高鹏
零九的思念
高鹏
号之前被盗了，请小心，连微信的好友都被人删了，还若无其事的用着，这什么心态啊！
22:43:07
问你个问题
考验下你的智商
哥是靠头脑吃饭的，难不倒我啦
为什么牛顿没吃掉从树上掉下来的苹果呢？
额…
消息记录
英特尔芯平板：无数好友点赞
关闭(C)
发送(S)
搜索
热点： 综艺 电视剧 礼物 礼包 游戏
苹果
苹果前任CEO解释苹果为何不在印度推…
来自腾讯科技
苹果无法为大众市场提供规格较低的产品，因为它的商业模式决定了它只能在世界上大多数国家销售高端智能手机，攫取…
iPhone 4
苹果印度停售iPhone 4 担心平均售价…
来自腾讯科技
在截至3月末的第二财季，iPhone的全球平均销售价格下滑41美元，降至596美元。
苹果跃居美国第二大网络零售商
来自腾讯科技
2013年，苹果网络销售额增长至183亿美元，增幅达24%。
苹果股价两年来首次站上600美元
来自腾讯科技
苹果股价曾在2012年9月份创下702.10美元的历史最高水平。
更多推荐： 苹果股价 苹果公司

Social media

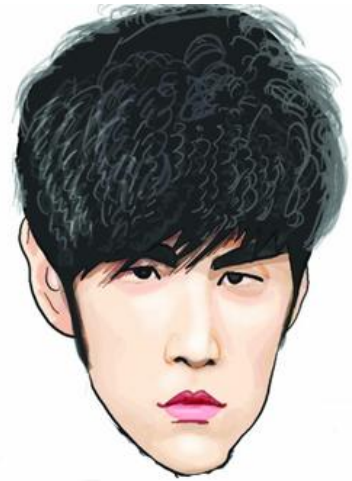


Our Approach



Chinese Family Name-based Binary Classification

In micro-blogs, if the front of to-be-tested entity has the sign '@', such entity does not exist in the knowledge base, 'NIL' will be returned straightforwardly.



Where is JieLun Zhou?

Chinese Family Name-based Binary Classification

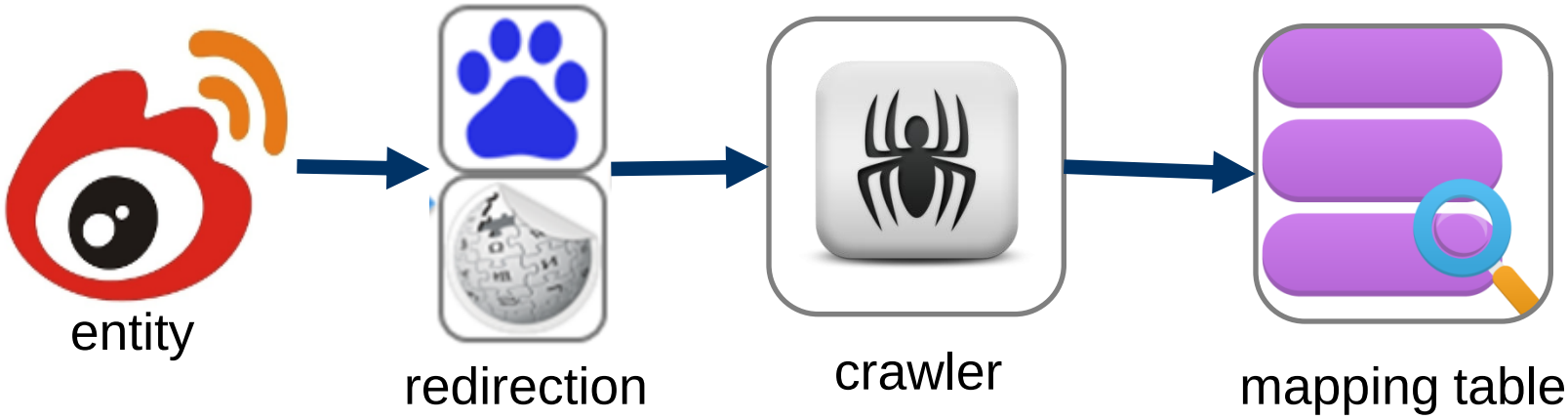
Xie TingFeng → Xiao Xie

Feng DeLun → Lao Feng

Considering the situation of putting the surname in the first, such as "Xie TingFeng", we will get the first word of the to-be-tested entity to lookup Chinese Family Name table. Considering the situation of putting the surname in the last, such as " Xiao Xie" , we will get the last word of the to-be-tested entity to lookup Chinese Family Name table.

Mapping Table

We use Baidu Encyclopedia and Wikipedia page redirection to get the mapped entity of to-be-tested entity.



Mapping Table

ShenZhou VI

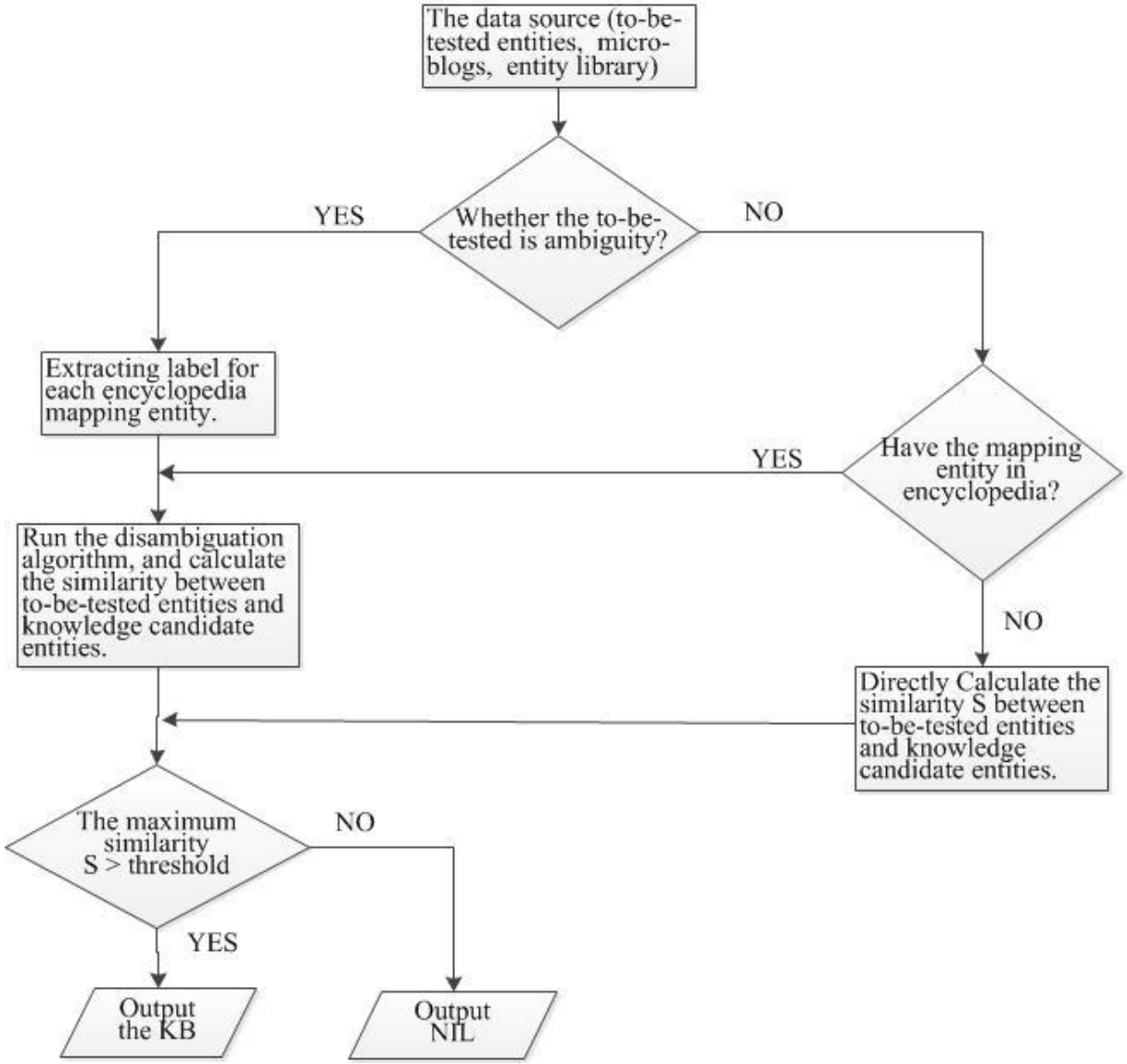
Baidu Encyclopedia page redirection

ShenZhou six spacecraft

Wikipedia page redirection

ShenZhou VI

Label Disambiguation



An improved label disambiguation algorithm

For Example:

We have a micro-blog, the to-be-tested entity is apple, and knowledge base has two different apples, Apple1 and Apple2.

Their attribute values are shown next.
Apple 1 attributes A= {Fruits, agricultural products, plants, trees, apple treesg}

Apple 2 attributes B = {Apple, Steve Jobs, electronic products, corporate, iPhone, Apple products}

An improved label disambiguation algorithm

The attribute of to-be-tested entity of apple are as follow:

C={Kumquat, fruit, stamps, China, postal services, agricultural products, Chenggu, Lintong, pomegranates, Luochuan, apples, liquan}

To-be-tested entity apple's space for text entry that doesn't duplicate set of

attributes as follows:

D={Kumquat, fruit, stamps, China, postal services, agricultural products, Chenggu, Lintong, pomegranates, Luochuan, apples, Liquan, plants, trees, apple trees, Apple, Steve Jobs, electronic products, corporate, iPhone, Apple products}

An improved label disambiguation algorithm

If the ambiguity item attributes appear in D counts as 1, if the relationship counts as 0.5, otherwise 0. We get the following similarity vectors.

$\vec{A}'$=(0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0.5, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0)

$\vec{B}'$=(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)

$\vec{C}'$=(1, 1, 1, 1, 1, 1 1, 1, 1, 1, 1, 1, 0 , 0, 0.5, 0, 0, 0, 0, 0, 0)

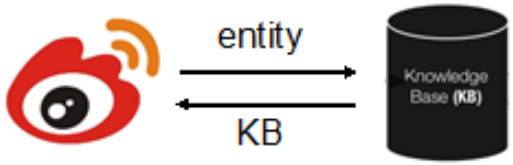
Put C' as a benchmark reference items, when the same position has the same value plus 1, when the same position has the value 1 and 0.5 plus 0.5, on the other case plus 0. So the similarity of A is M1 = 10, and the similarity of B is M2 = 2.5, thus the to-be-tested-entity links to Apple 1.

An improved label disambiguation algorithm

Assuming Pi is the value of benchmark reference items at position i, Qi is the value of control items at position i. The calculating formula of similarity degree can be described as follow:

$$SIM(\mathrm{k}) = \begin{cases} k+1, \{(P_i, Q_i) \mid P_i = Q_i\} \\ k+0.5, \{(P_i, Q_i) \mid P_i \in S, Q_i \in S \text{且} P_i \neq Q_i\} \end{cases}$$

Experimental Results and Analysis



Experimental results and analysis

These three combination are:

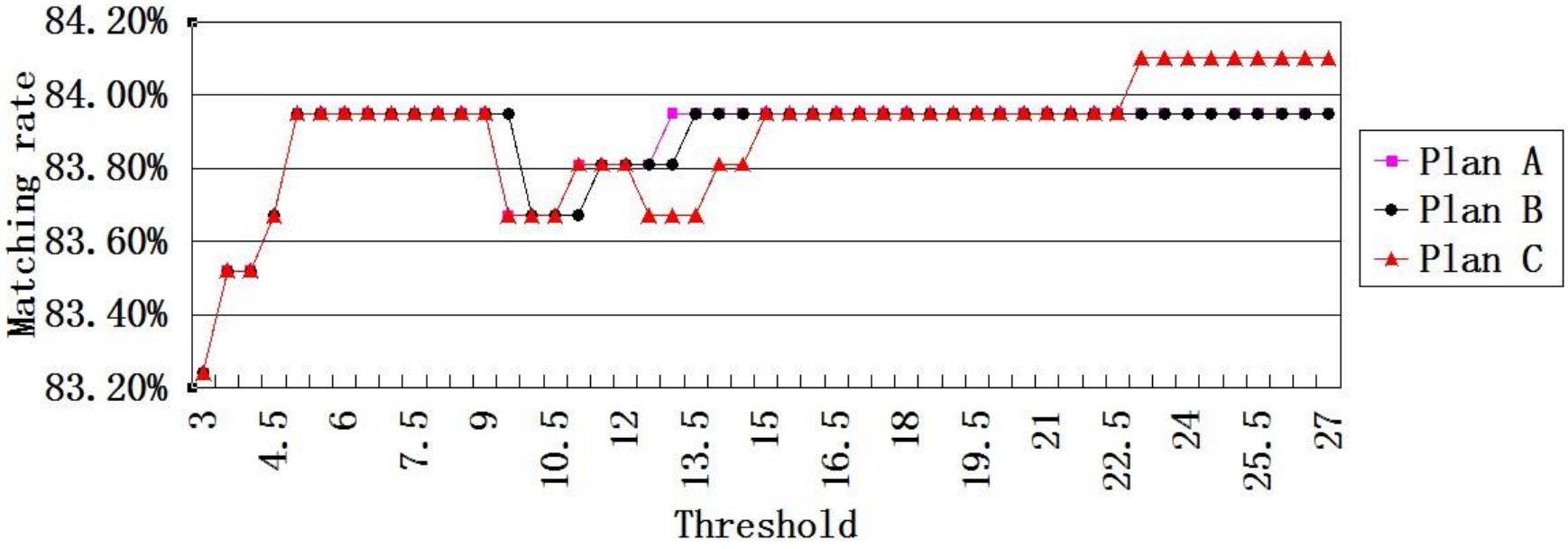
Plan A: look up Chinese Family Name table, Baidu mapping table, Wiki mapping table, direct matching and label disambiguation algorithm.

Plan B: look up Chinese Family Name table, Baidu mapping table, Wiki mapping table, label disambiguation algorithm and direct matching.

Plan C: look up Chinese Family Name table, direct matching, Baidu mapping table, Wiki mapping table and label disambiguation algorithm.

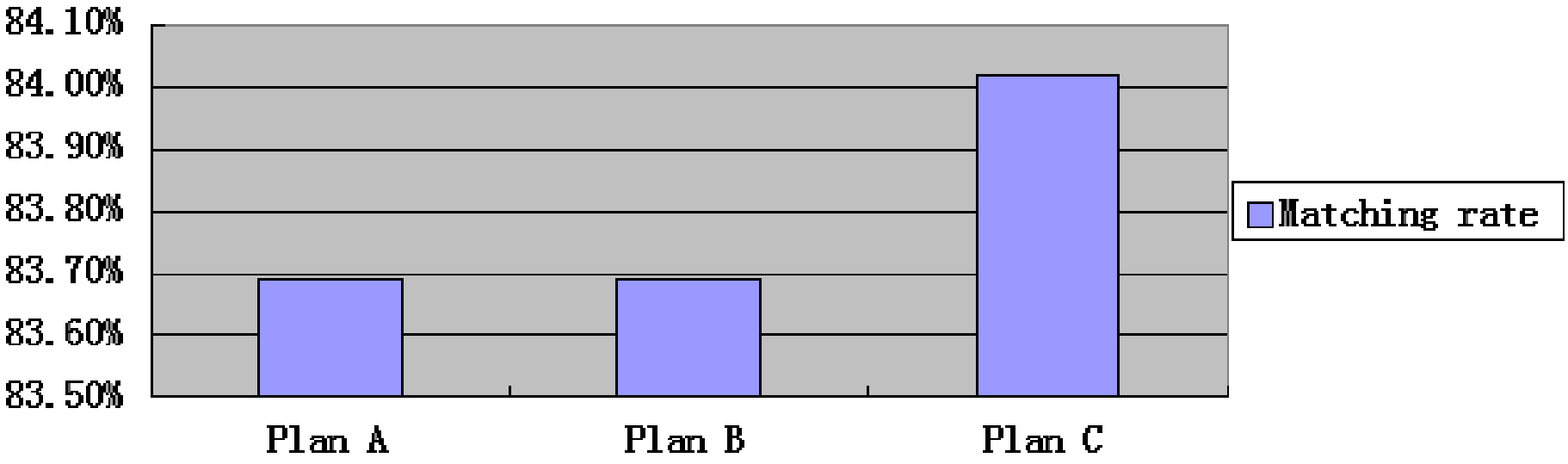
Thresholds and Matching Rate

Threshold and matching rates



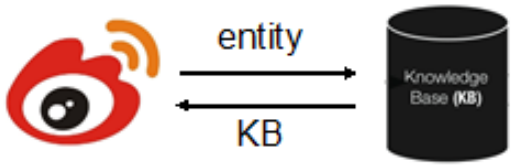
Optimal Strategy

The relationships between different strategies and matching rates



Plan C: look up Chinese Family Name table, direct matching, Baidu mapping table, Wiki mapping table and label disambiguation algorithm.

Conclusion



Evaluation Results

Chinese micro-blog entity link evaluation results are shown in Table .

Evaluation	Best	Overall		In-KB		NIL		Ranking
average	evaluation	Correct#	Accuracy	Precision	Recall	Precision	Recall	Third
70.58%	86.82%	510	0.8402	0.8103	0.7765	0.8640	0.8892	

Table 1: Evaluation results

Thanks!
Q&A?