

# CASIA at WSC2008

---

Institute of Automation  
Chinese Academy of Sciences

Guanggang Geng

[guanggang.geng@ia.ac.cn](mailto:guanggang.geng@ia.ac.cn)

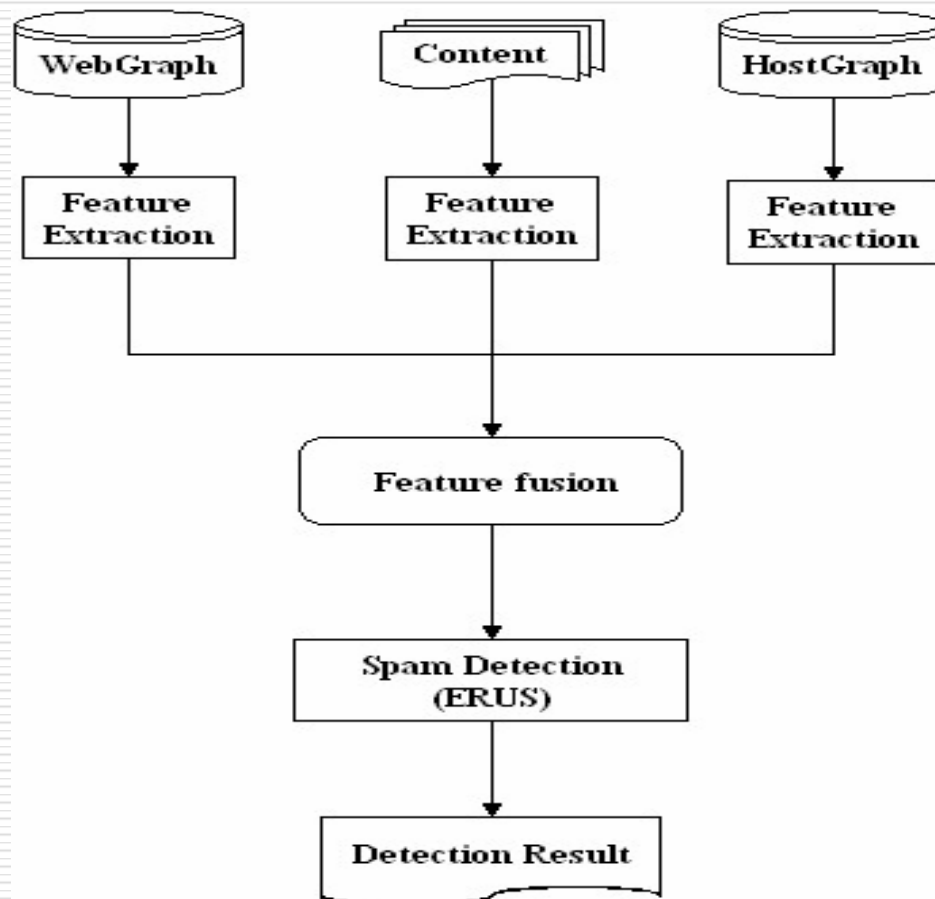
Xiaobo Jin

[xbjin@nlpr.ia.ac.cn](mailto:xbjin@nlpr.ia.ac.cn)

Chunheng Wang

[chunheng.wang@ia.ac.cn](mailto:chunheng.wang@ia.ac.cn)

# Detection Framework



# Host Level link analysis Features

$$F_1(h) = Measure(h)$$

$$F_2(h) = \sum_{v \in Inlink(h)} Measure(v) * weight(v, h)$$

$$F_3(h) = \sum_{v \in Outlink(h)} Measure(v) * weight(h, v)$$

$$F_4(h) = \frac{\sum_{v \in Inlink(h)} Measure(v) * weight(v, h)}{\sum_{v \in Inlink(h)} weight(v, h)}$$

$$F_5(h) = \frac{\sum_{v \in Outlink(h)} Measure(v) * weight(h, v)}{\sum_{v \in Outlink(h)} weight(h, v)}$$

*Measures???*

*HostRank,*

*TrustRank,*

*Truncated  
PageRank (TP)  
(T=1,2..K)*

*weight(h,v)=f(n), n is the  
the number of hyperlinks  
from host h to host v*

*we use boolean weight*

# Host Level link analysis Features

$$F_6(h) = \frac{\sum_{v \in \text{Inlink}(\text{Inlink}(h))} \text{Measure}(v)}{|\text{Inlink}(\text{Inlink}(h))|}$$

$$F_7(h) = \frac{\sum_{v \in \text{Inlink}(\text{Outlink}(h))} \text{Measure}(v)}{|\text{Inlink}(\text{Outlink}(h))|}$$

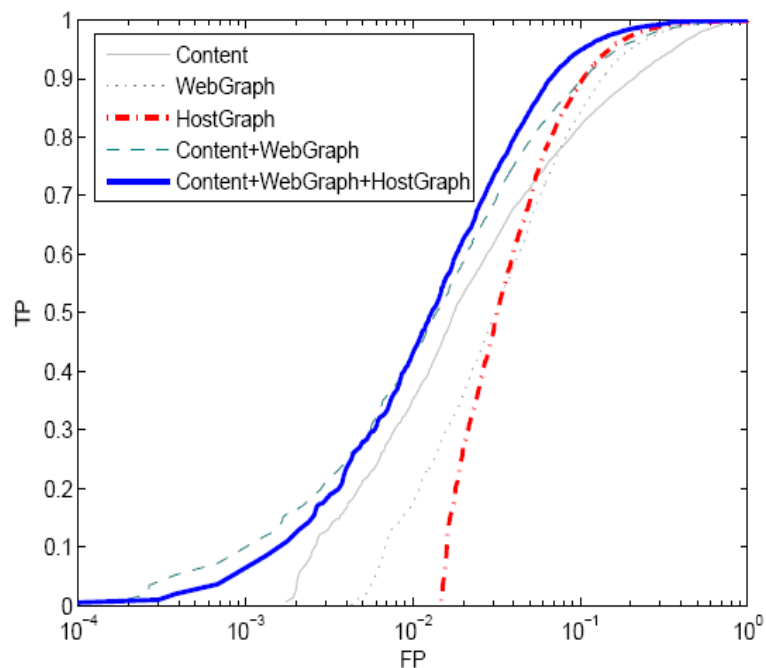
$$F_8(h) = \frac{\sum_{v \in \text{Outlink}(\text{Inlink}(h))} \text{Measure}(v)}{|\text{Outlink}(\text{Inlink}(h))|}$$

$$F_9(h) = \frac{\sum_{v \in \text{Outlink}(\text{Outlink}(h))} \text{Measure}(v)}{|\text{Outlink}(\text{outlink}(h))|}$$

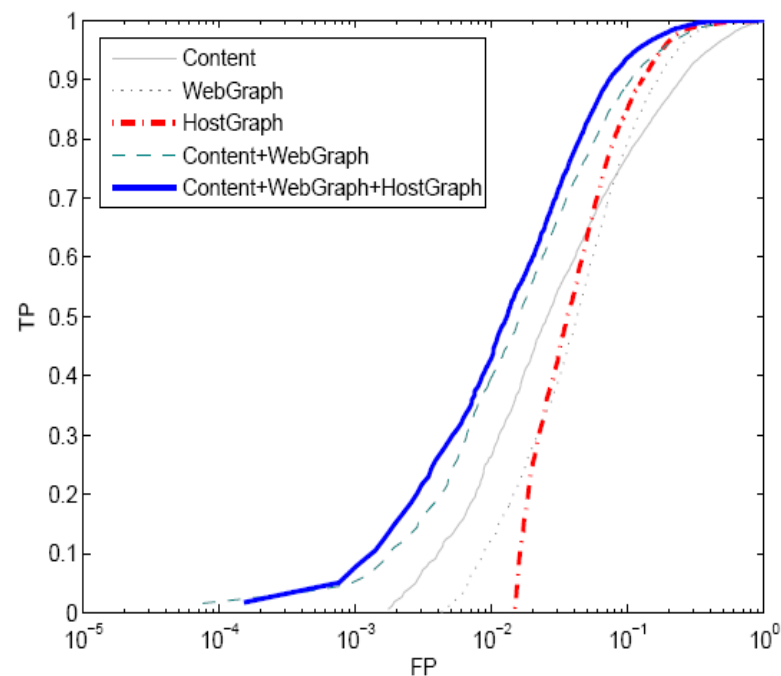
$$F_{10}(h) = \text{SiteSupporter}_d(h) \quad d \in \{1, 2, \dots, k\}$$

We extract  $9 * 4(\text{HostRank, trustRank, TP (T=1,2)}) + 4(d=1,2,3,4) = 40$  host level link features

# Performance with different features on **WEBSPAM-UK2006 (Set1 + Set2)** (5-CV)



ROC with Bagging(C4.5)



ROC with Adaboost(stump)

# Performance with different features on **WEBSPAM-UK2006 (Set1 + Set2)** (5-CV)

Features	Precision	Recall	F1-measure	AUC
Content(C)	0.807	0.712	0.756	0.915
WebGraph(W)	0.771	0.781	0.776	0.931
HostGraph(H)	0.775	0.857	0.814	0.941
C+W	0.839	0.828	0.833	0.959
C+W+H	0.852	0.873	0.862	0.969

Bagging(C4.5)

Features	Precision	Recall	F1-measure	AUC
Content(C)	0.839	0.740	0.786	0.931
WebGraph(W)	0.793	0.816	0.804	0.942
HostGraph(H)	0.805	0.860	0.831	0.949
C+W	0.845	0.832	0.838	0.960
C+W+H	0.855	0.887	0.871	0.971

Adaboost(stump)

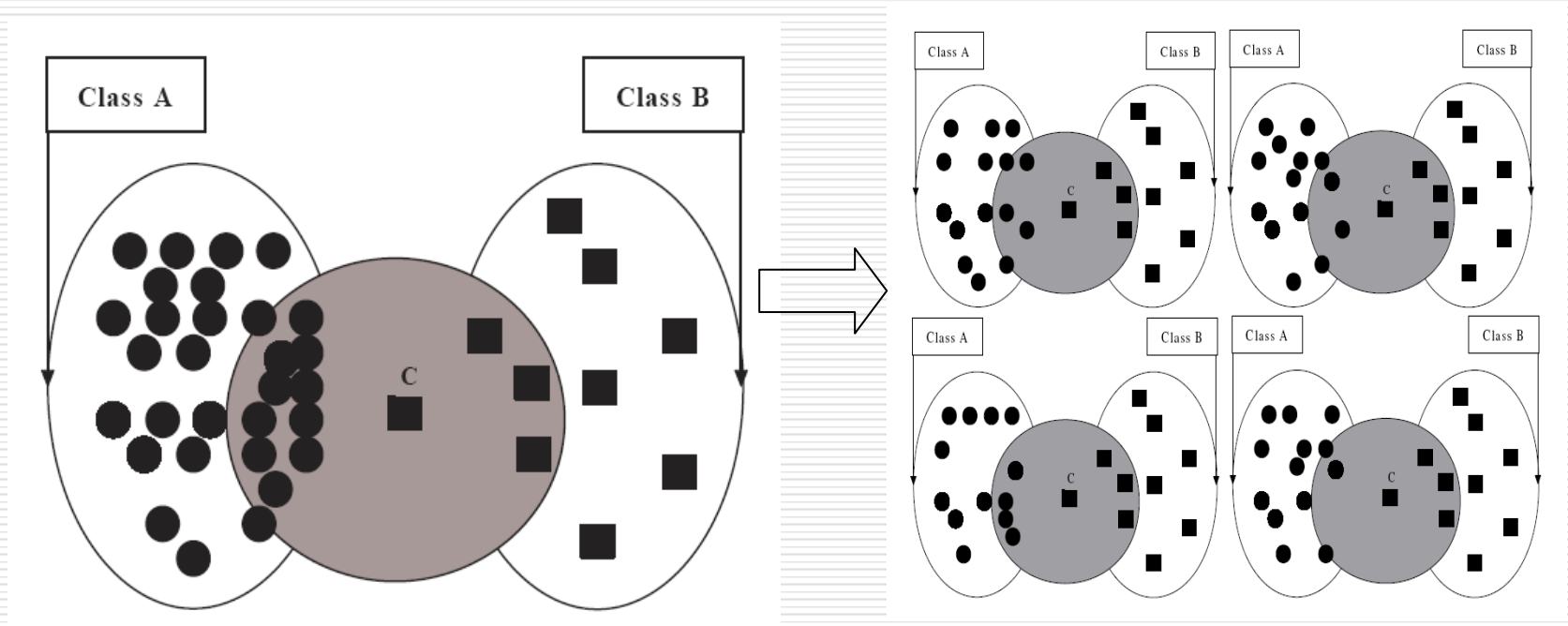
# Detection Strategy---Ensemble Random Under-Sampling(ERUS)

How imbalance???

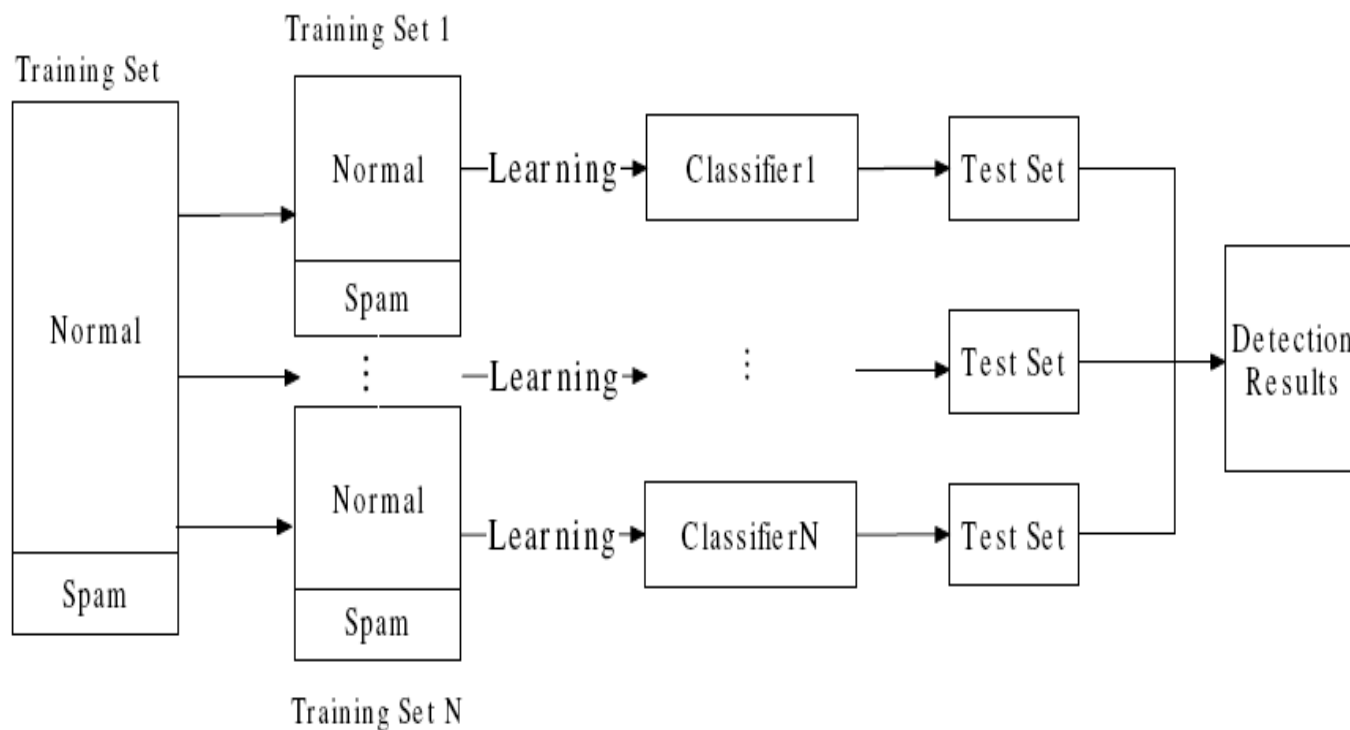
Non-spam: Spam

WEBSpAM-UK2006-Set1 7:1

WEBSpAM-UK2007-Set1 18:1



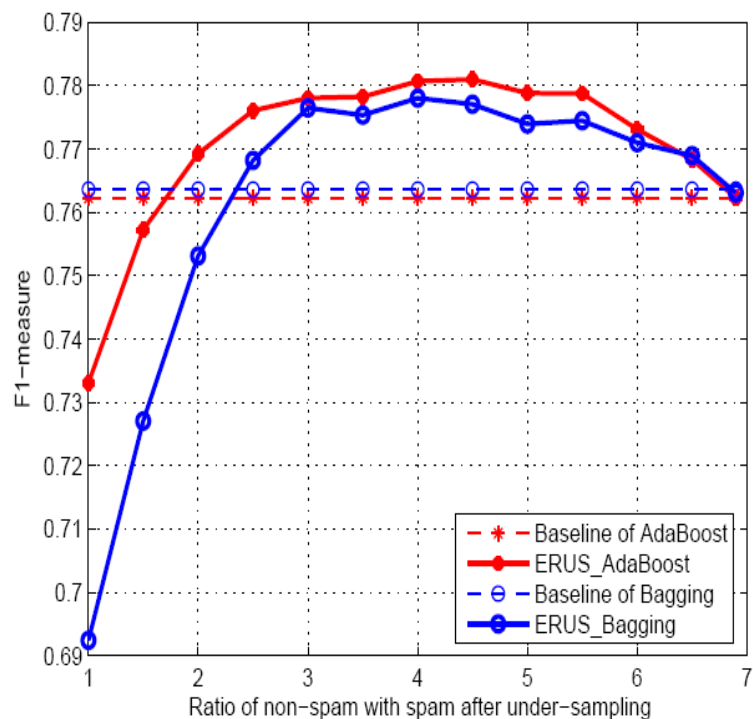
# Detection Strategy---Ensemble Random Under-Sampling(ERUS)



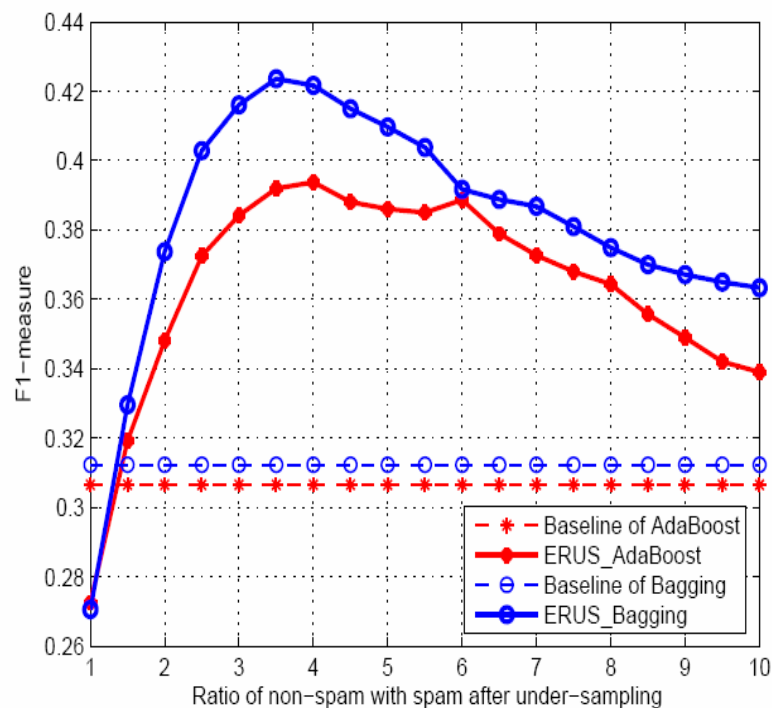
*Integration is based on the predicted spamicity of all the resampled training sets. In our experiment, the resample times  $T=15$*



# ERUS Performance---F1-Measure (5-CV)

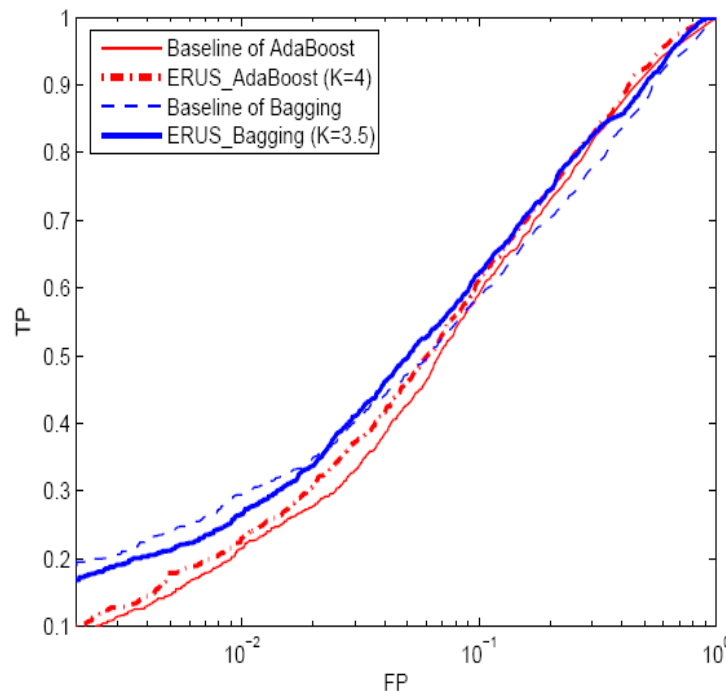
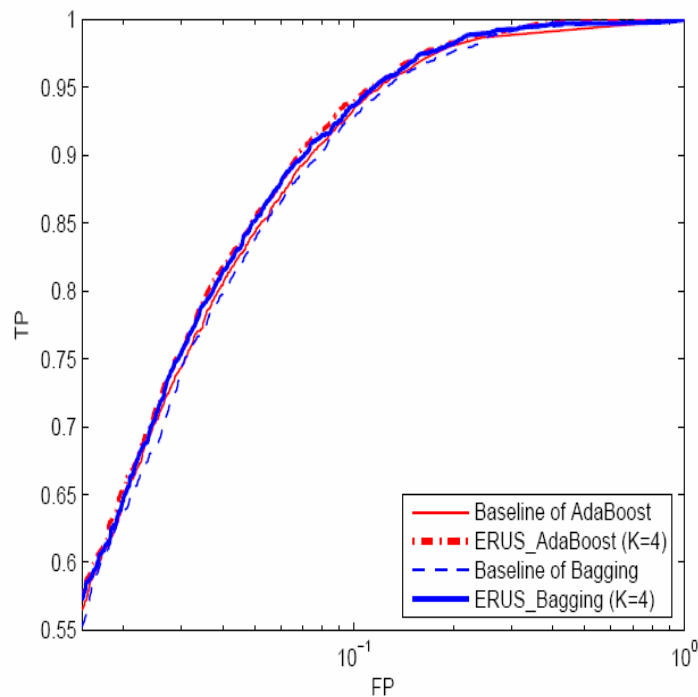


WEBSpam-UK2006-Set1



WEBSpam-UK2007-Set1

# ERUS Performance---ROC (5-CV)



$K = |\text{non-spam}| / |\text{spam}|$ , (after under-sampling)

WEBSPAM-UK2006-Set1

WEBSPAM-UK2007-Set1

# ERUS Performance (5-CV)

WEBSPAM-UK2006-SET1

Measures	AdaBoost	ERUS_Ada	Bagging	ERUS_Bag
F1-measure	0.762	<b>0.781</b>	0.763	0.778
AUC	0.967	<b>0.972</b>	0.968	<b>0.972</b>

WEBSPAM-UK2007-SET1

Measures	AdaBoost	ERUS_Ada	Bagging	ERUS_Bag
F1-measure	0.307	0.394	0.312	<b>0.424</b>
AUC	0.841	<b>0.855</b>	0.829	0.851

---

# Thank you!