MIS Quarterly

RESEARCH ARTICLE

EXTRACTING REPRESENTATIVE INFORMATION ON INTRA-ORGANIZATIONAL BLOGGING PLATFORMS

Xunhua Guo, Qiang Wei, and Guoqing Chen
Research Center for Contemporary Management, School of Economics and Management, Tsinghua University,
Beijing 100084 CHINA {guoxh@sem.tsinghua.edu.cn} {weiq@sem.tsinghua.edu.cn}
{chengq@sem.tsinghua.edu.cn}

Jin Zhang
School of Business, Renmin University of China,
Beijing 100872 CHINA {zhangjin@rbs.ruc.edu.cn}

Dandan Qiao
Research Center for Contemporary Management, School of Economics and Management, Tsinghua University,
Beijing 100084 CHINA {qiaodd.12@sem.tsinghua.edu.cn}

Appendix A

An Illustrative Example for the Clustering Process of REPSET

Figure A1 illustrates the clustering process of the REPSET method. There are nine documents represented as five round points and four square points. The distances between data points correspond to the pairwise similarities between documents. Intuitively, these nine documents can be divided into two clusters, i.e., the round point cluster and the square point cluster.

At the beginning, each point is treated as an individual cluster. In each step, two clusters with the largest similarity are merged into a new one. The backward strategy is applied after the generation of a new cluster. In the example, in steps (b)–(f), no boundary document is found in the generated cluster and reallocation does not occur. In step (g), two documents located in the middle are merged into one cluster, since they have the largest similarity. If the clustering process continues as the traditional hierarchical clustering method, the nine documents will finally be grouped into two clusters as shown in step (h), which turns out to be inaccurate, because one square document is assigned to a cluster in which most documents are round. In contrast, in REPSET, the square document will be marked as a boundary object (outlined in red), since its average similarity to the whole cluster is less than λ. The backward strategy will then reevaluate the red document and reallocate it if needed. As shown in (i), because the red document's similarity to the left cluster is than that to the right cluster, it is reallocated to the left cluster.

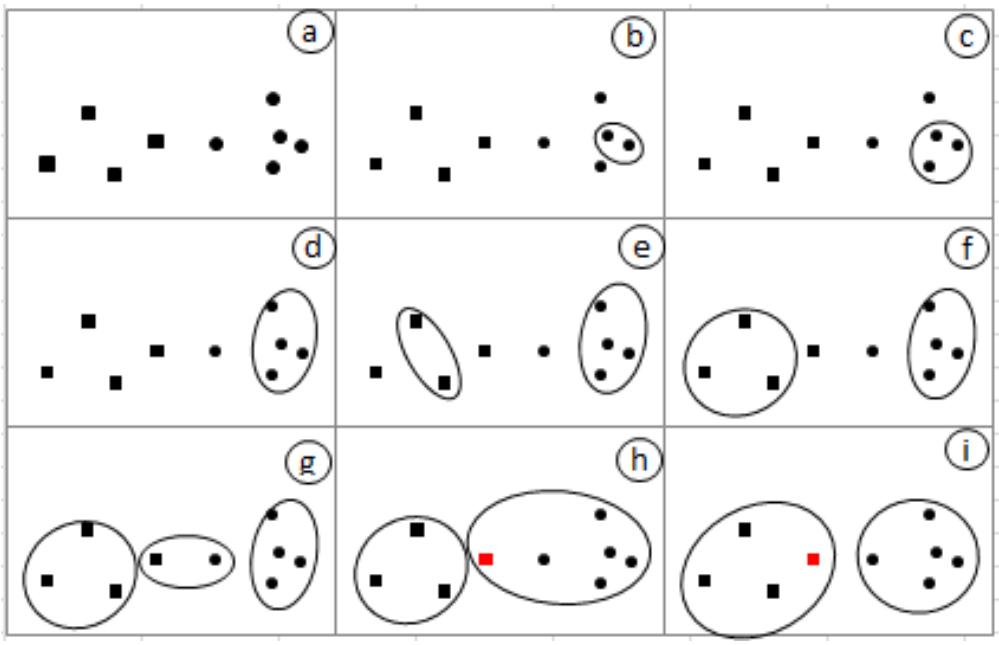


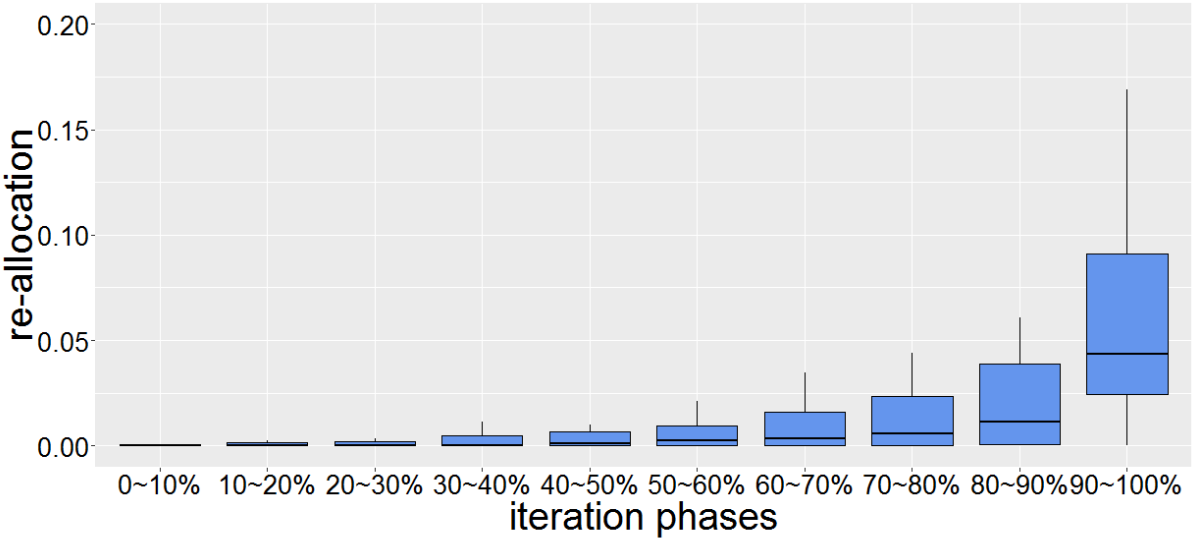
Figure A1. Illustrative Example for REPSET Clustering

Appendix B

Proportions of Data Objects Reallocated in REPSET Clustering

Table B1: Proportions of Data Objects Reallocated in REPSET Clustering

Dataset No.	Reallocation Proportion	Dataset No.	Reallocation Proportion
1	0.91%	16	3.23%
2	0.15%	17	1.17%
3	0.00%	18	0.37%
4	1.35%	19	0.25%
5	0.01%	20	5.79%
6	4.42%	21	1.52%
7	2.16%	22	1.16%
8	0.00%	23	0.04%
9	3.47%	24	0.78%
10	0.35%	25	1.08%
11	0.38%	26	1.16%
12	0.51%	27	0.60%
13	0.47%	28	0.18%
14	4.94%	29	0.00%
15	0.30%	30	2.37%
Average			1.29%



Note: In each iteration phases, the endpoints of each line represent the maximum and minimum re-allocation values respectively, and the three lines of each box represent the upper quartile, the median, and the lower quartile respectively from top to bottom.

Figure B1. Proportions of Data Objects Reallocated in Iteration Stages

Appendix C

Representative Article Extraction Example

As an example, we compare the results of representative article extraction between REPSET and X-Means, a typical traditional clustering method. We used the blogging data of a typical region of Company X, during the time period of July 2010, with a total of 2,239 blog articles. Using REPSET and X-Means, respectively, we extracted 10 articles from the data set. Figure C1 shows the comparison of the sizes of clusters generated by the two methods. It can be seen that the sizes of clusters generated by X-Means is much more even than those of REPSET. These results illustrate the fact that X-Means tends to select clusters that are spatially equidistant, which is deficient for representativeness extraction. In contrast, the sizes of clusters generated by REPSET are more discrepant, indicating the capability of REPSET to capture more diverse content.

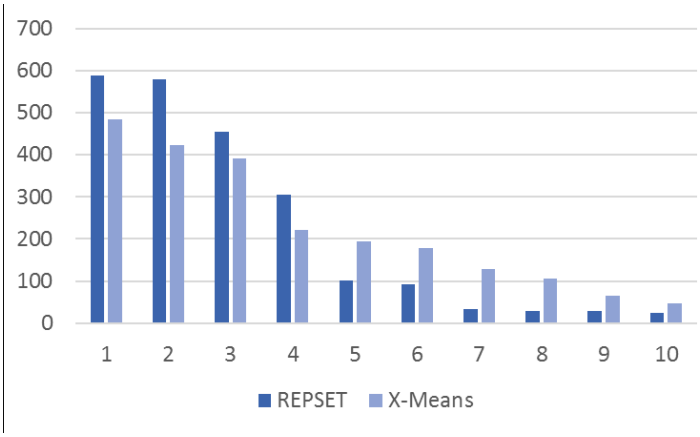


Figure C1. Cluster Size Comparison

Table C1 lists the titles of the representative articles extracted by REPSET and X-Means, respectively. It can be seen that the articles extracted by X-Means are mostly life-related and emotional essays. In the whole data set, about 70% of articles are life-related and 30% are work-related. In the results of X-Means, work-related articles are drown in life-related content and fail to be revealed. Meanwhile, most of the extracted life-related articles are similar to some extent. In contrast, in the results of REPSET, six extracted articles are work-related, covering diverse topics including work balance, industry analysis, and customer alerts. The four life-related articles are also diverse, covering topics such as love, dining, and life advices. Such results intuitively show that REPSET can facilitate enhanced diversity in coverage and therefore capture more representative results.

Table C1. Representative Article Extraction Results

Articles Extracted by REPSET

No.	Article Title	Category
1	In the south, love is not only conveyed with poetry	Life
2	Nice places for dining in the Shunde city	Life
3	How to balance project work and post work	Work
4	Fourteen advices for girls born in the 1980s	Life
5	Industry Watch: China Mobile is facing five challenges in the field of information technology	Work
6	The most comprehensive methods for recovering from drunk	Life
7	Sharing the morning business meeting memo for July 6, 2010	Work
8	Training plan for new employees	Work
9	Work plan, 2nd week, July, Ronggui customer services	Work
10	VIP customer alert numbers	Work

Articles Extracted by X-Means

No.	Article Title	Category
1	Happiness blind	Life
2	There is a love that cannot be waited, there is a love that cannot be hurt.	Life
3	Life is like donkeys, dogs, and monkeys.	Life
4	Workplace is like a battlefield	Work
5	To lose	Life-work
6	Sadness is a kind of beauty	Life
7	Love is only one more stroke than hate	Life
8	Someone to allow me to be unreasonable	Life
9	Losing affection	Life
10	Spider nets	Life

Appendix D

Empirical Data Experiments Results with Varying Numbers of Representative Articles

CLUSTER NUM	MEASURE	REPSET	RANDOM	TOP RATED	MOST READ	MOST COMMENTED	GRAPH
10	de facto F_1-measure	0.504	0.391	0.278	0.374	0.205	0.326
10	de facto Coverage	0.374	0.284	0.193	0.287	0.130	0.326
10	de facto Redundancy	0.227	0.373	0.502	0.463	0.521	0.675
15	de facto F_1-measure	0.442	0.4	0.286	0.348	0.212	0.356
15	de facto Coverage	0.34	0.331	0.219	0.311	0.146	0.382
15	de facto Redundancy	0.371	0.492	0.588	0.606	0.619	0.667
20	de facto F_1-measure	0.444	0.396	0.274	0.333	0.211	0.368
20	de facto Coverage	0.379	0.357	0.235	0.341	0.162	0.409
20	de facto Redundancy	0.465	0.554	0.671	0.675	0.697	0.665
25	de facto F_1-measure	0.458	0.391	0.258	0.313	0.206	0.382
25	de facto Coverage	0.437	0.373	0.246	0.353	0.171	0.457
25	de facto Redundancy	0.519	0.588	0.73	0.718	0.741	0.672
30	de facto F_1-measure	0.438	0.371	0.245	0.295	0.194	0.359
30	de facto Coverage	0.445	0.4	0.271	0.36	0.177	0.477
30	de facto Redundancy	0.569	0.653	0.776	0.75	0.785	0.712
35	de facto F_1-measure	0.433	0.349	0.239	0.27	0.183	0.338
35	de facto Coverage	0.458	0.407	0.281	0.369	0.182	0.492
35	de facto Redundancy	0.59	0.694	0.792	0.787	0.817	0.742
40	de facto F_1-measure	0.401	0.349	0.261	0.249	0.177	0.337
40	de facto Coverage	0.466	0.427	0.373	0.382	0.188	0.51
40	de facto Redundancy	0.648	0.705	0.799	0.815	0.832	0.748
45	de facto F_1-measure	0.387	0.33	0.245	0.238	0.174	0.318
45	de facto Coverage	0.459	0.44	0.379	0.389	0.196	0.52
45	de facto Redundancy	0.665	0.735	0.819	0.828	0.844	0.771
50	de facto F_1-measure	0.378	0.312	0.233	0.223	0.164	0.307
50	de facto Coverage	0.465	0.442	0.384	0.394	0.2	0.53
50	de facto Redundancy	0.682	0.76	0.833	0.844	0.861	0.784
55	de facto F_1-measure	0.371	0.299	0.216	0.217	0.214	0.296
55	de facto Coverage	0.474	0.456	0.392	0.407	0.336	0.534
55	de facto Redundancy	0.696	0.778	0.851	0.852	0.843	0.795
60	de facto F_1-measure	0.357	0.302	0.215	0.209	0.211	0.279
60	de facto Coverage	0.483	0.456	0.397	0.413	0.356	0.536
60	de facto Redundancy	0.717	0.775	0.853	0.86	0.85	0.812
70	de facto F_1-measure	0.33	0.268	0.193	0.187	0.198	0.251
70	de facto Coverage	0.521	0.483	0.408	0.426	0.382	0.549
70	de facto Redundancy	0.759	0.815	0.873	0.88	0.866	0.837
80	de facto F_1-measure	0.311	0.251	0.199	0.17	0.192	0.227
80	de facto Coverage	0.533	0.49	0.415	0.433	0.403	0.557

CLUSTER NUM	MEASURE	REPSET	RANDOM	TOP RATED	MOST READ	MOST COMMENTED	GRAPH
80	de facto Redundancy	0.78	0.831	0.869	0.894	0.874	0.857
90	de facto F_1-measure	0.306	0.238	0.197	0.155	0.176	0.212
90	de facto Coverage	0.54	0.499	0.425	0.44	0.416	0.563
90	de facto Redundancy	0.787	0.844	0.872	0.906	0.888	0.87
100	de facto F_1-measure	0.285	0.231	0.187	0.143	0.168	0.192
100	de facto Coverage	0.547	0.51	0.433	0.448	0.426	0.565
100	de facto Redundancy	0.808	0.851	0.881	0.915	0.895	0.885
110	de facto F_1-measure	0.276	0.215	0.172	0.135	0.174	0.195
110	de facto Coverage	0.555	0.52	0.437	0.454	0.433	0.572
110	de facto Redundancy	0.817	0.864	0.893	0.921	0.891	0.882
120	de facto F_1-measure	0.25	0.192	0.162	0.151	0.163	0.178
120	de facto Coverage	0.559	0.528	0.445	0.459	0.44	0.577
120	de facto Redundancy	0.839	0.883	0.901	0.91	0.9	0.895
130	de facto F_1-measure	0.243	0.193	0.16	0.144	0.152	0.174
130	de facto Coverage	0.563	0.535	0.453	0.465	0.445	0.584
130	de facto Redundancy	0.845	0.882	0.903	0.914	0.908	0.898
140	de facto F_1-measure	0.225	0.187	0.154	0.138	0.143	0.163
140	de facto Coverage	0.57	0.543	0.463	0.473	0.453	0.59
140	de facto Redundancy	0.86	0.887	0.907	0.919	0.915	0.906
150	de facto F_1-measure	0.215	0.171	0.146	0.135	0.137	0.153
150	de facto Coverage	0.575	0.551	0.467	0.478	0.459	0.592
150	de facto Redundancy	0.868	0.899	0.913	0.921	0.92	0.912
CLUSTER NUM	**MEASURE**	**XMEANS**	**RBR**	**DIRECT**	**LDA**	**HLDA**	**DTM**
10	de facto F_1-measure	0.433	0.469	0.452	0.36	0.409	0.37
10	de facto Coverage	0.365	0.402	0.4	0.284	0.311	0.272
10	de facto Redundancy	0.468	0.437	0.48	0.51	0.405	0.423
15	de facto F_1-measure	0.424	0.474	0.467	0.339	0.395	0.369
15	de facto Coverage	0.391	0.438	0.431	0.317	0.324	0.28
15	de facto Redundancy	0.536	0.483	0.491	0.635	0.497	0.456
20	de facto F_1-measure	0.396	0.464	0.467	0.294	0.379	0.394
20	de facto Coverage	0.421	0.463	0.463	0.326	0.34	0.353
20	de facto Redundancy	0.627	0.536	0.529	0.732	0.572	0.553
25	de facto F_1-measure	0.364	0.436	0.419	0.286	0.359	0.385
25	de facto Coverage	0.434	0.483	0.486	0.337	0.352	0.363
25	de facto Redundancy	0.686	0.603	0.632	0.752	0.632	0.591
30	de facto F_1-measure	0.336	0.41	0.43	0.287	0.348	0.429
30	de facto Coverage	0.446	0.497	0.495	0.356	0.365	0.391
30	de facto Redundancy	0.731	0.652	0.621	0.76	0.668	0.525
35	de facto F_1-measure	0.322	0.382	0.389	0.258	0.334	0.42
35	de facto Coverage	0.47	0.511	0.508	0.363	0.374	0.406
35	de facto Redundancy	0.755	0.696	0.685	0.8	0.698	0.565
40	de facto F_1-measure	0.305	0.38	0.372	0.252	0.32	0.389
40	de facto Coverage	0.478	0.521	0.518	0.38	0.381	0.383

CLUSTER NUM	MEASURE	XMEANS	RBR	DIRECT	LDA	HLDA	DTM
40	de facto Redundancy	0.776	0.701	0.71	0.812	0.725	0.604
45	de facto F_1-measure	0.287	0.349	0.371	0.249	0.296	0.396
45	de facto Coverage	0.488	0.531	0.529	0.456	0.391	0.424
45	de facto Redundancy	0.797	0.74	0.715	0.829	0.762	0.63
50	de facto F_1-measure	0.268	0.359	0.369	0.205	0.277	0.337
50	de facto Coverage	0.494	0.538	0.537	0.386	0.398	0.449
50	de facto Redundancy	0.816	0.73	0.719	0.86	0.788	0.73
55	de facto F_1-measure	0.259	0.342	0.346	0.206	0.259	0.357
55	de facto Coverage	0.503	0.545	0.545	0.396	0.399	0.451
55	de facto Redundancy	0.825	0.751	0.747	0.861	0.809	0.705
60	de facto F_1-measure	0.254	0.333	0.356	0.179	0.254	0.352
60	de facto Coverage	0.505	0.554	0.552	0.396	0.413	0.471
60	de facto Redundancy	0.83	0.762	0.737	0.884	0.817	0.72
70	de facto F_1-measure	0.23	0.304	0.338	0.182	0.267	0.354
70	de facto Coverage	0.532	0.564	0.562	0.488	0.437	0.483
70	de facto Redundancy	0.853	0.792	0.758	0.888	0.808	0.72
80	de facto F_1-measure	0.213	0.296	0.307	0.165	0.245	0.292
80	de facto Coverage	0.546	0.574	0.57	0.493	0.444	0.49
80	de facto Redundancy	0.868	0.801	0.79	0.901	0.831	0.792
90	de facto F_1-measure	0.194	0.288	0.301	0.146	0.233	0.301
90	de facto Coverage	0.558	0.583	0.582	0.506	0.452	0.51
90	de facto Redundancy	0.883	0.809	0.797	0.914	0.843	0.787
100	de facto F_1-measure	0.183	0.263	0.283	0.139	0.219	0.303
100	de facto Coverage	0.565	0.592	0.588	0.513	0.455	0.494
100	de facto Redundancy	0.891	0.831	0.814	0.92	0.856	0.781
110	de facto F_1-measure	0.174	0.269	0.27	0.132	0.206	0.247
110	de facto Coverage	0.572	0.6	0.598	0.52	0.46	0.519
110	de facto Redundancy	0.898	0.827	0.826	0.924	0.867	0.838
120	de facto F_1-measure	0.162	0.244	0.244	0.121	0.194	0.244
120	de facto Coverage	0.58	0.607	0.605	0.523	0.465	0.535
120	de facto Redundancy	0.906	0.847	0.847	0.932	0.877	0.842
130	de facto F_1-measure	0.156	0.225	0.239	0.105	0.185	0.232
130	de facto Coverage	0.581	0.613	0.61	0.452	0.47	0.519
130	de facto Redundancy	0.91	0.863	0.852	0.941	0.885	0.851
140	de facto F_1-measure	0.145	0.228	0.233	0.102	0.172	0.215
140	de facto Coverage	0.586	0.621	0.617	0.461	0.473	0.545
140	de facto Redundancy	0.918	0.861	0.856	0.943	0.895	0.866
150	de facto F_1-measure	0.133	0.207	0.229	0.098	0.163	0.231
150	de facto Coverage	0.589	0.625	0.623	0.54	0.476	0.54
150	de facto Redundancy	0.925	0.876	0.86	0.946	0.901	0.853

Appendix E

Experiment Script (Translated from Chinese)

Experiments on Blog Article Reading

NO: _____

Dear students,

Welcome to participate in this experiment!

The following are ten articles selected from a company's internal blogging system. Please carefully read these blog articles for 20 minutes. When the time is up, you will be asked to answer several questions mark a number of words according to what you have read.

Basic information:

1. Name_____, Gender_____, Age_____
2. Major_____, Grade_____
3. Your familiarity with blogging system _____
   A. Completely unfamiliar B. Unfamiliar C. Neutral D. Familiar E. Very familiar

(After providing the basic information, the articles are displayed and the subjects are required to read the articles for 20 minutes. When the time is up, the system closes the article and presents the following questions.)

For each of the following question, please select the number that corresponds with your reading experience. The number "1" represents "strongly disagree," while the number "7" represents "strongly agree."

Questions	Please select						
	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
This group of blog articles has much repetition.	1	2	3	4	5	6	7
This group of blog articles has similar topics.	1	2	3	4	5	6	7
This group of blog articles has a wide range of contents.	1	2	3	4	5	6	7
This group of blog articles provides rich information.	1	2	3	4	5	6	7

If you have any feelings, opinions or comments on this group of blog articles please feel free to write them down in the following space:

Please label the following words based on your reading.

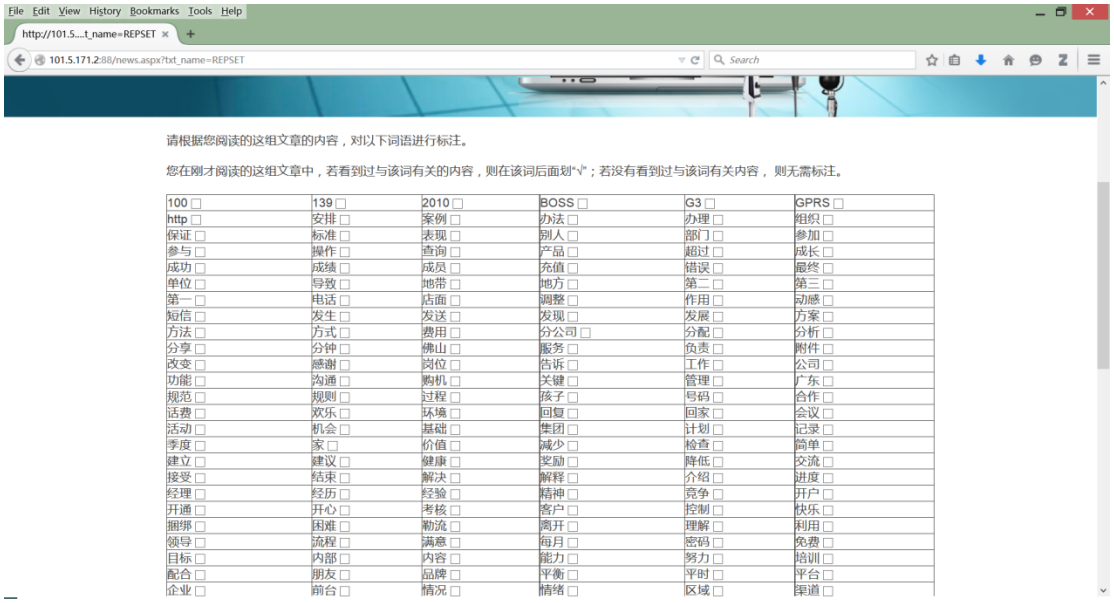
If you have browsed any information related with the word, please tick "√", otherwise no label is needed.

100	139	2010	BOSS	G3	GPRS
http	arrangement	case	method	transaction	organization
warranty	standard	performance	others	department	participate
take part in	operation	inquire	product	surpass	growth
success	grade	member	charge	error	final
unit	cause	zone	place	second	third
first	phone	store	adjustment	effect	dynamic
sms	happen	send	discovery	development	program
method	manner	fee	branch	allocation	analysis
share	minute	Foshan	service	responsible	accessory
change	thanks	post	tell	the work	company
function	communication	purchase	key	management	Guangdong
specification	rule	process	kids	number	cooperation
bill	joy	environment	reply	home	conference
activity	opportunity	foundation	group	plan	record
quarter	home	value	reduce	inspection	simple
set up	proposal	health	reward	drop	exchange
accept	end	solve	explanation	introduction	progress
manager	experience	experience	spirit	competition	account
opening	happy	check	client	control	happiness
binding	difficulty	leliu	leave	understanding	utilize
leadership	process	satisfaction	monthly	password	free
goal	internal	content	ability	strive	train
coordinate	friend	brand	balance	usually	platform
business	reception	situation	mood	area	channel
global pass	life	staff	daily	Ronggui	social
application	close	identity	body	life	life
time	world	market	thing	receive	collect
cell phone	familiar	data	Shunde	attitude	discussion
package	put forward	improve	provide	upgrade	remind
experience	condition	call	communicate	communications	notification
colleague	unity	complaints	team	spread	recommendation
expand	network	website	future	hope	habit
love	system	afternoon	show	on site	relevant
enjoy	project	consumption	sales	effect	assist
thanks	mood	mentality	information	week	industry
form	happiness	demand	propaganda	choose	student
learning	pressure	business	opinion	awareness	factor
marketing	operate	affect	own	forever	user
preferential	outstanding	pre-store	staff	reason	operations
online	responsibility	increase	gift	about	correct
policy	support	knowledge	execution	index	guide
formulate	quality	intelligence	china	center	terminal
emphasis	initiative	theme	status	consultation	charges
data	resources	comprehensive			

Appendix F

Screenshots for the Experiment System UI


File Edit View History Bookmarks Tools Help
http://101.5....t_name=REPSET
101.5.171.2:88/news.aspx?txt_name=REPSET
Search
文章阅读/news
伍佳网络正在改变
感谢您的关注，我们一直在成长
请根据您阅读的感受，对以下问题按1-7分打分，1为"非常不同意"，7为"非常同意"：
问题
请选择
非常不同意 不同意 比较不同意 不确定 比较同意 同意 非常同意
1. 这组文章的内容有很多重复。 1 2 3 4 5 6 7
2. 这组文章的话题相近。 1 2 3 4 5 6 7
3. 这组文章的内容范围广泛。 1 2 3 4 5 6 7
4. 这组文章提供了丰富的信息。 1 2 3 4 5 6 7


File Edit View History Bookmarks Tools Help
http://101.5....t_name=REPSET
101.5.171.2:88/news.aspx?txt_name=REPSET
Search
请根据您阅读的这组文章的内容，对以下词语进行标注。
您在刚才阅读的这组文章中，若看到过与该词有关的内容，则在该词后面划"√"；若没有看到过与该词有关内容，则无需标注。
100 139 2010 BOSS G3 GPRS
http 安排 案例 办法 办理 组织
保证 标准 表现 别人 部门 参加
参与 操作 查询 产品 超过 成长
成功 成绩 成员 充值 错误 最终
单位 导致 地带 地方 第二 第三
第一 电话 店面 调整 作用 动感
短信 发生 发送 发现 发展 方案
方法 方式 费用 分公司 分配 分析
分享 分钟 佛山 服务 负责 附件
改变 感谢 岗位 告诉 工作 公司
功能 沟通 购机 关键 管理 广东
规范 规则 过程 孩子 号码 合作
话费 欢乐 环境 回复 回家 会议
活动 机会 基础 集团 计划 记录
季度 家 价值 减少 检查 简单
建立 建议 健康 奖励 降低 交流
接受 结束 解决 解释 介绍 进度
经理 经历 经验 精神 竞争 开户
开通 开心 考核 客户 控制 快乐
捆绑 困难 勒流 离开 理解 利用
领导 流程 满意 每月 密码 免费
目标 内部 内容 能力 努力 培训
配合 朋友 品牌 平衡 平时 平台
企业 前台 情况 情绪 区域 渠道