マスコミュニケーション学会発表

キーワード分析による
発信者別WWWコンテンツ量の推計

２００１年１０月８日

郵政研究所　島田　博也

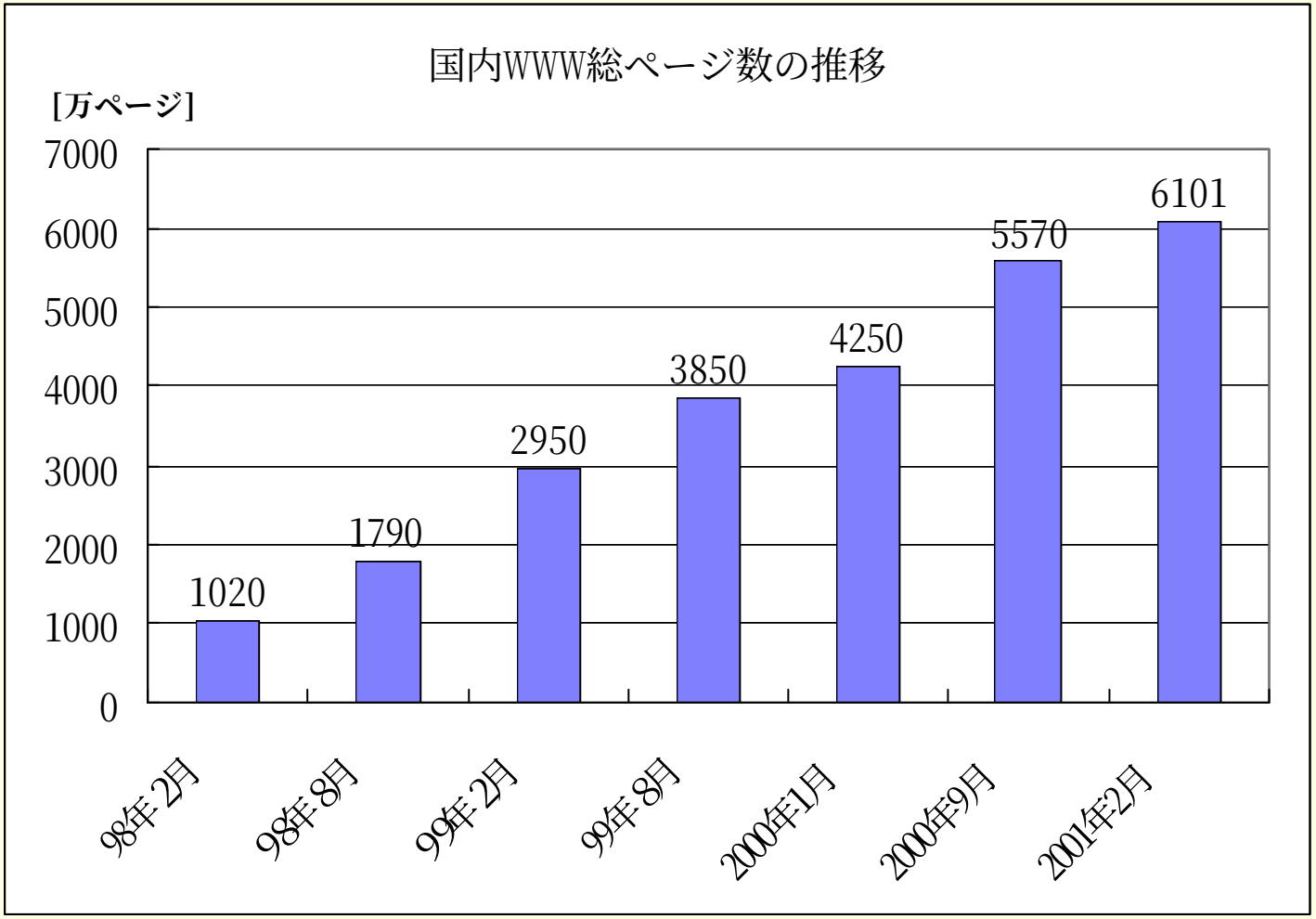
アライド・ブレインズ株式会社　内田　斉

東京大学大学院　大岩　寛

1　WWWの急成長

Allied Brains

- □　WWWの総ページ数は６０００万ページ以上と推定される。
- □　これは、新聞・雑誌が年間に提供する情報量をしのぐ。
- □　３年間で６倍の規模に成長し、巨大な「マスメディア」に。

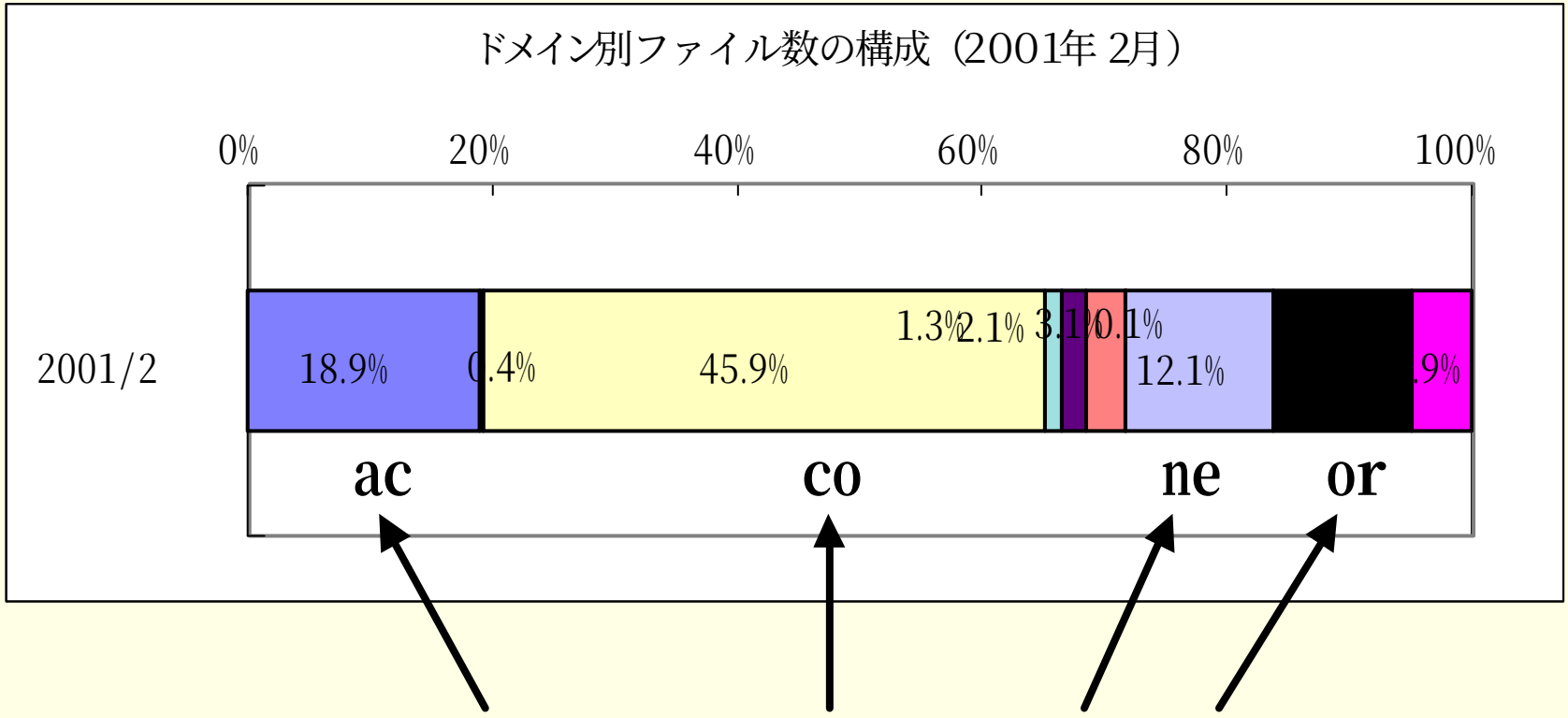


2　WWWの特徴としての個人ホームページ

Allied Brains

□　**WWWが他のマスメディアと決定的に異なるのは「個人が自由に情報発信できる」という点。**

□　**しかし、個人ホームページの総量はまったく不明。**



個人ホームページに該当するドメインは存在しない。

３　個人ホームページの識別方法

Allied Brains

□　ドメイン名やＵＲＬから個人ホームページを識別することはできない。
□　個人ホームページでしか用いられない言葉があれば、その言葉が含まれているかどうかによって個人ホームページを識別することが考えられる。

識別キーワードの条件

ある特定のカテゴリーのサイトでのみ出現率が高く、
それ以外のサイトではほとんど出現しない言葉。

考えられる例

個人ＨＰ・・・ぼく、あたし、家族、息子、娘、日記・・・など
学校ＨＰ・・・本校、校歌、卒業、進路、生徒・・・など

問題点

キーワードによりページの発信者を１００％識別することは不可能。
特に個人ホームページはテーマが雑多なので難しい。

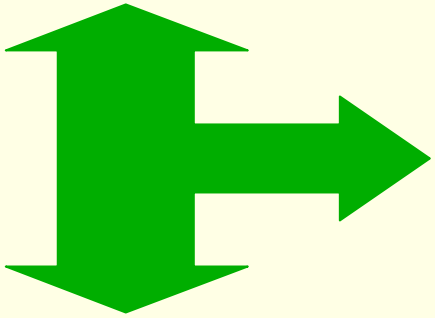
４　個人ホームページ総量の推計方法

Allied Brains

□　すべての個人ホームページが特定できなくても、総量の推計は可能。
□　識別キーワードの「個人ホームページにおける出現率」と「ウェブ全体での出現率」が分かれば、ウェブ全体での個人ホームページの構成比を推定できる。

例：「家族紹介」という単語が、個人ホームページのみに出現する言葉で、
出現率の調査結果が以下のようであった場合・・・

個人ホームページであることが分かっているページでの出現率：３０％



ウェブ全体での出現率：１０％

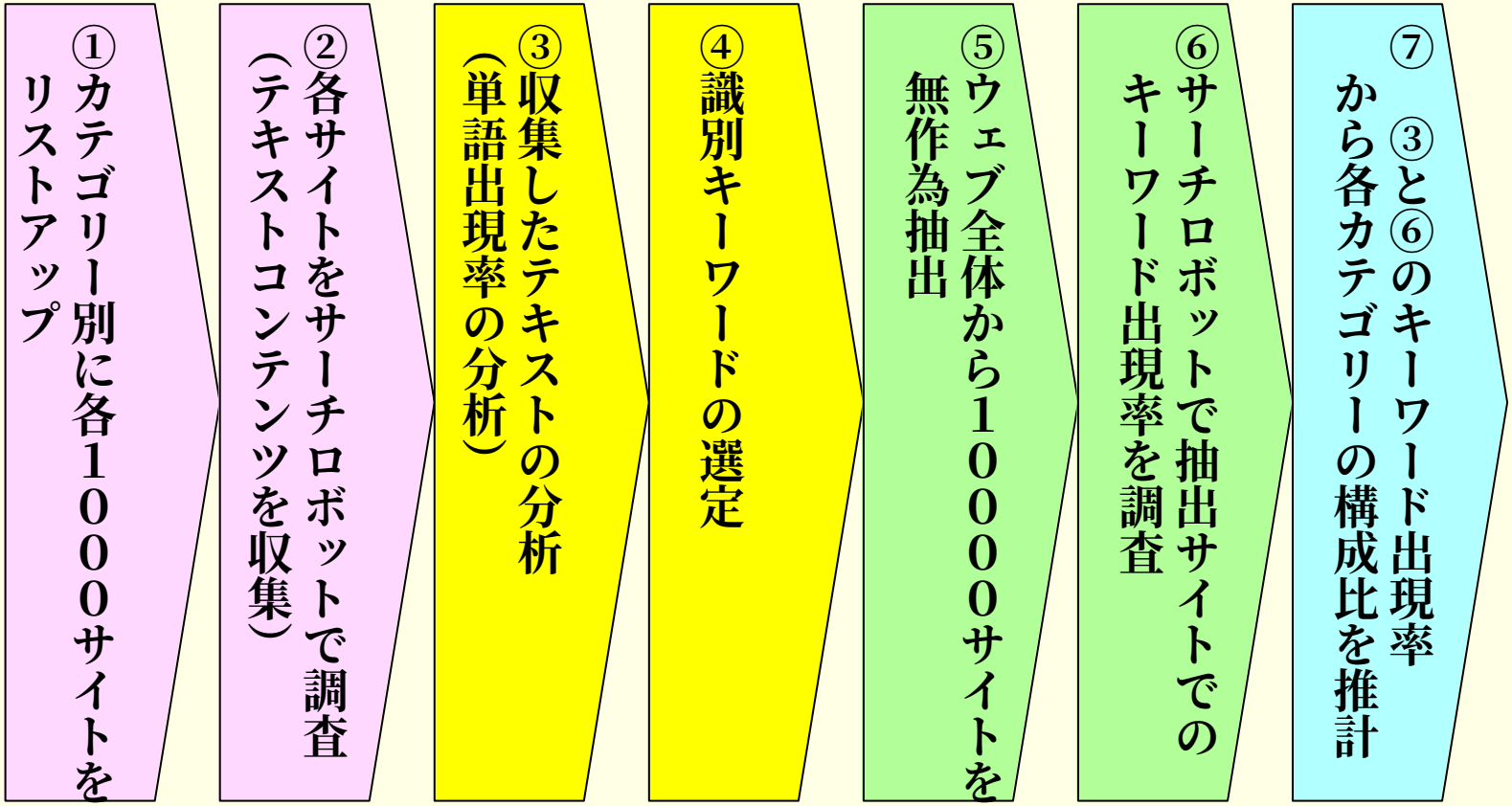
個人ホームページはウェブ全体の１／３を占めると推定できる。

5　調査の手順

Allied Brains

□　４つの発信者カテゴリー（個人・企業・自治体・学校）別に各１０００サイトを調査し、識別キーワード群を抽出。

□　次に、ウェブ全体から１００００サイトをランダム抽出し、識別キーワードの出現率を調査。



6　抽出した識別キーワード群

Allied Brains

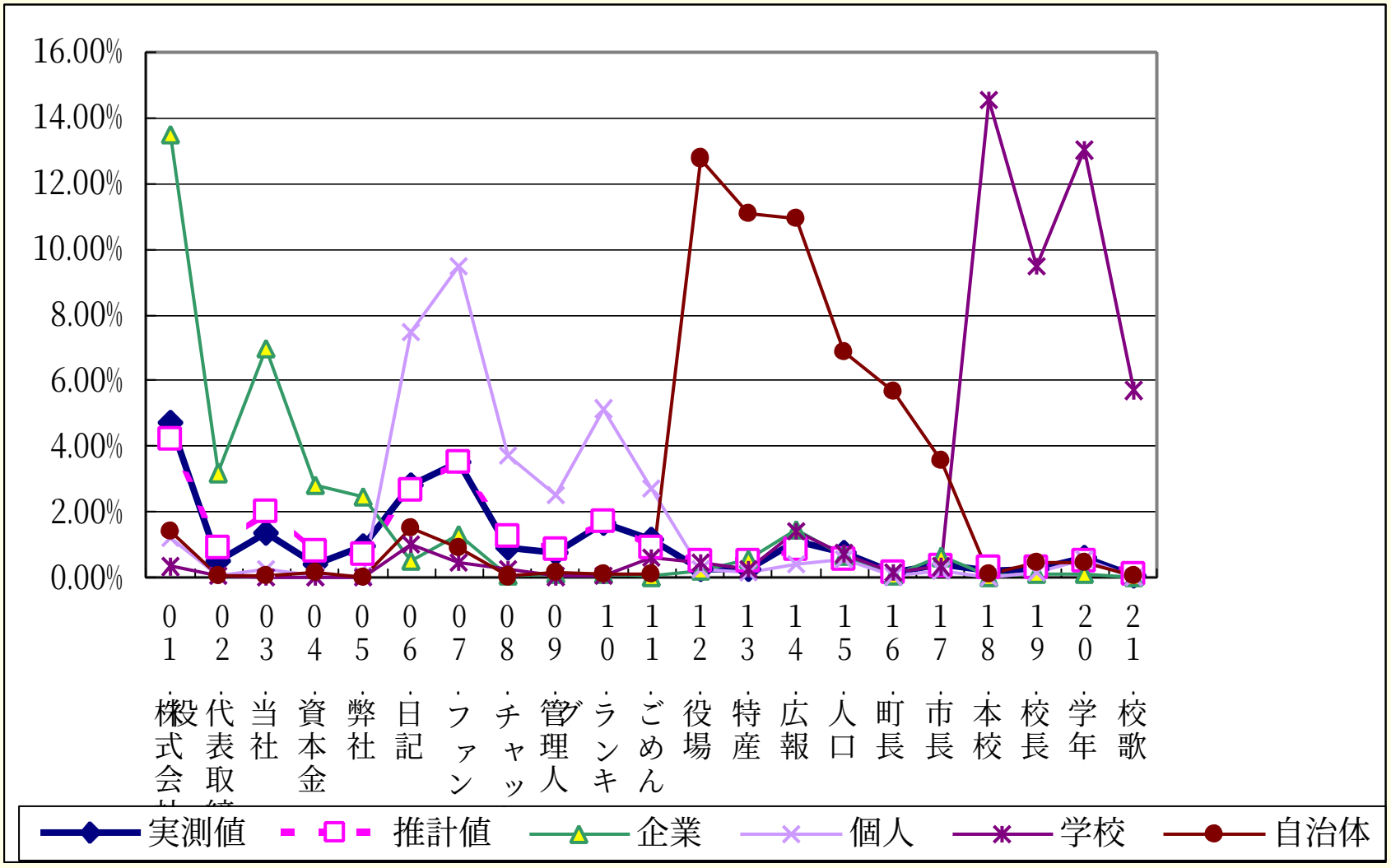
□　特定カテゴリーでのみ出現率が高い、21個の識別キーワードを選定。

		企業	個人	自治体	学校
1	株式会社	13.47%	1.23%	1.40%	0.33%
2	代表取締役	3.20%	0.05%	0.06%	0.05%
3	当社	6.96%	0.27%	0.06%	0.00%
4	資本金	2.81%	0.04%	0.15%	0.01%
5	弊社	2.48%	0.05%	0.01%	0.00%
6	日記	0.53%	7.47%	1.51%	1.00%
7	ファン	1.30%	9.50%	0.92%	0.47%
8	チャット	0.05%	3.70%	0.03%	0.24%
9	管理人	0.11%	2.51%	0.15%	0.03%
10	ランキング	0.13%	5.15%	0.13%	0.06%
11	ごめん	0.03%	2.72%	0.12%	0.60%
12	役場	0.20%	0.27%	12.76%	0.48%
13	特産	0.55%	0.16%	11.10%	0.24%
14	広報	1.44%	0.39%	10.92%	1.43%
15	人口	0.69%	0.55%	6.86%	0.74%
16	町長	0.06%	0.04%	5.67%	0.18%
17	市長	0.64%	0.20%	3.56%	0.33%
18	本校	0.03%	0.01%	0.10%	14.55%
19	校長	0.11%	0.16%	0.46%	9.50%
20	学年	0.13%	0.55%	0.45%	13.02%
21	校歌	0.00%	0.00%	0.07%	5.70%

7　識別キーワード群のウェブ全体での出現率

Allied Brains

□　ランダム抽出した１００００サイトでの、識別キーワードの出現率は下のグラフのとおり。



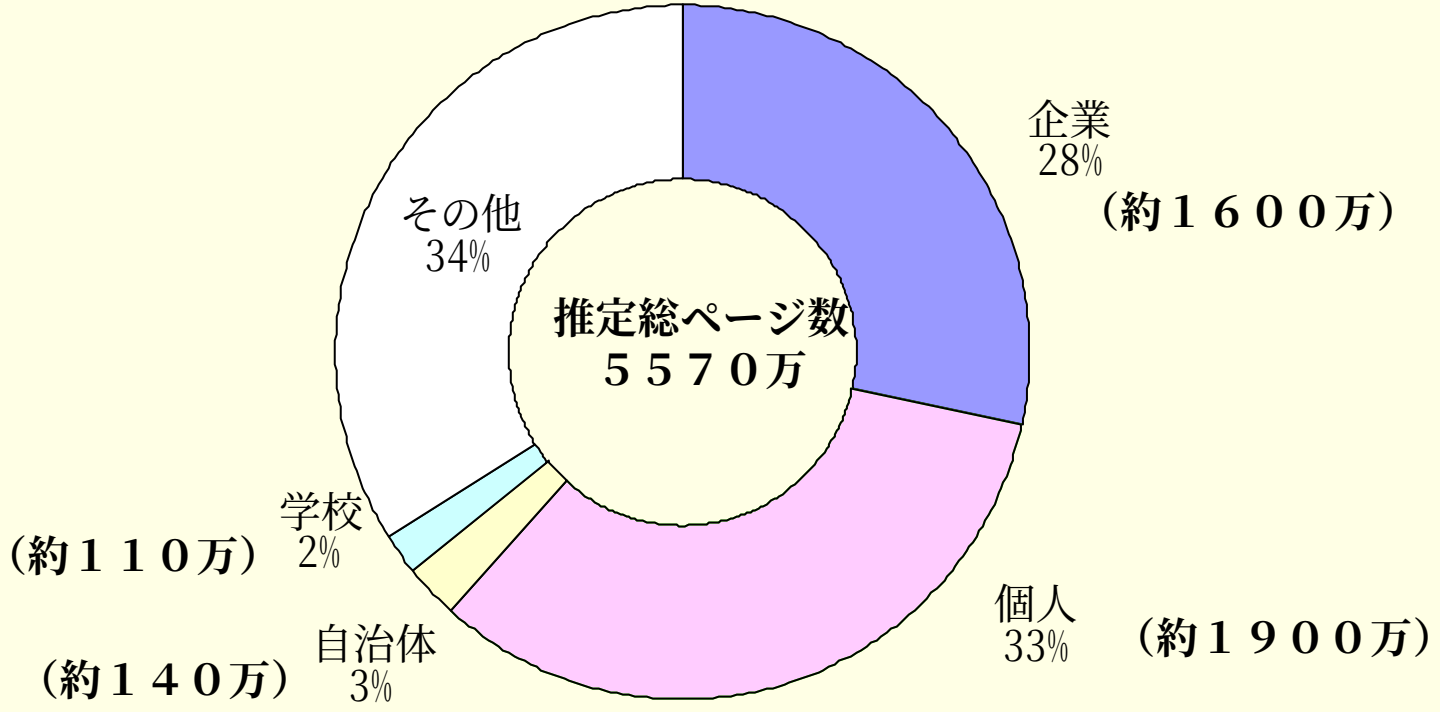
7　発信者カテゴリー別の総ページ数の推計結果

Allied Brains

□　重回帰分析により、ウェブ全体での各カテゴリーの構成比を推計。
□　個人ホームページは全体の約3割を占め、企業ホームページよりも若干多いと推定される。

回帰統計	
重相関 R	0.973095
重決定 R2	0.946914
補正 R2	0.878723
標準誤差	0.003017
観測数	21

	係数	標準誤差	t	P-値	下限 95%	上限 95%
	0	#N/A	#N/A	#N/A	#N/A	#N/A
企業	0.282033	0.019161	14.7193	4.18E-11	0.241608	0.322459
個人	0.333638	0.021654	15.40796	2.02E-11	0.287953	0.379323
自治体	0.025548	0.013715	1.862807	0.079862	-0.00339	0.054484
学校	0.020729	0.01347	1.538875	0.14224	-0.00769	0.049148

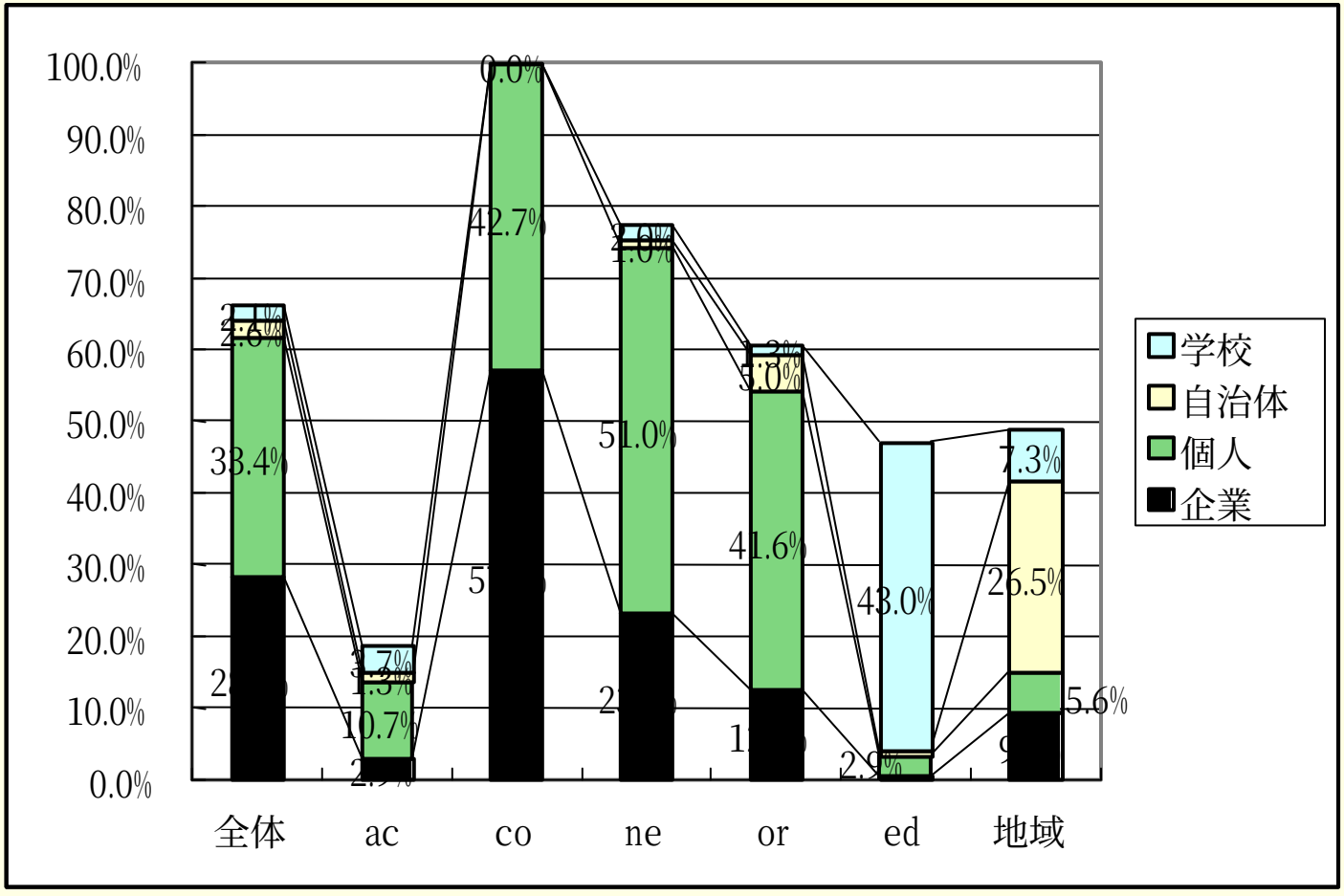


8　ドメインと発信者との関係

□　ドメイン別のキーワード出現率から、各カテゴリーの発信者がどのドメインでホームページを発信しているかを分析した。

□　個人ホームページは、複数のドメインに分布。

各ドメインは、どのカテゴリーの発信者のページで構成されているか



Allied Brains

１）これまでまったく不明だった、個人ホームページ数、発信者別ホームページ数の把握が可能に。

２）個人ホームページは、ウェブの中心的なコンテンツであることが定量的に確認できた。

３）汎用JPドメインの増加が予想される中で、ウェブ統計調査の手法として有効。