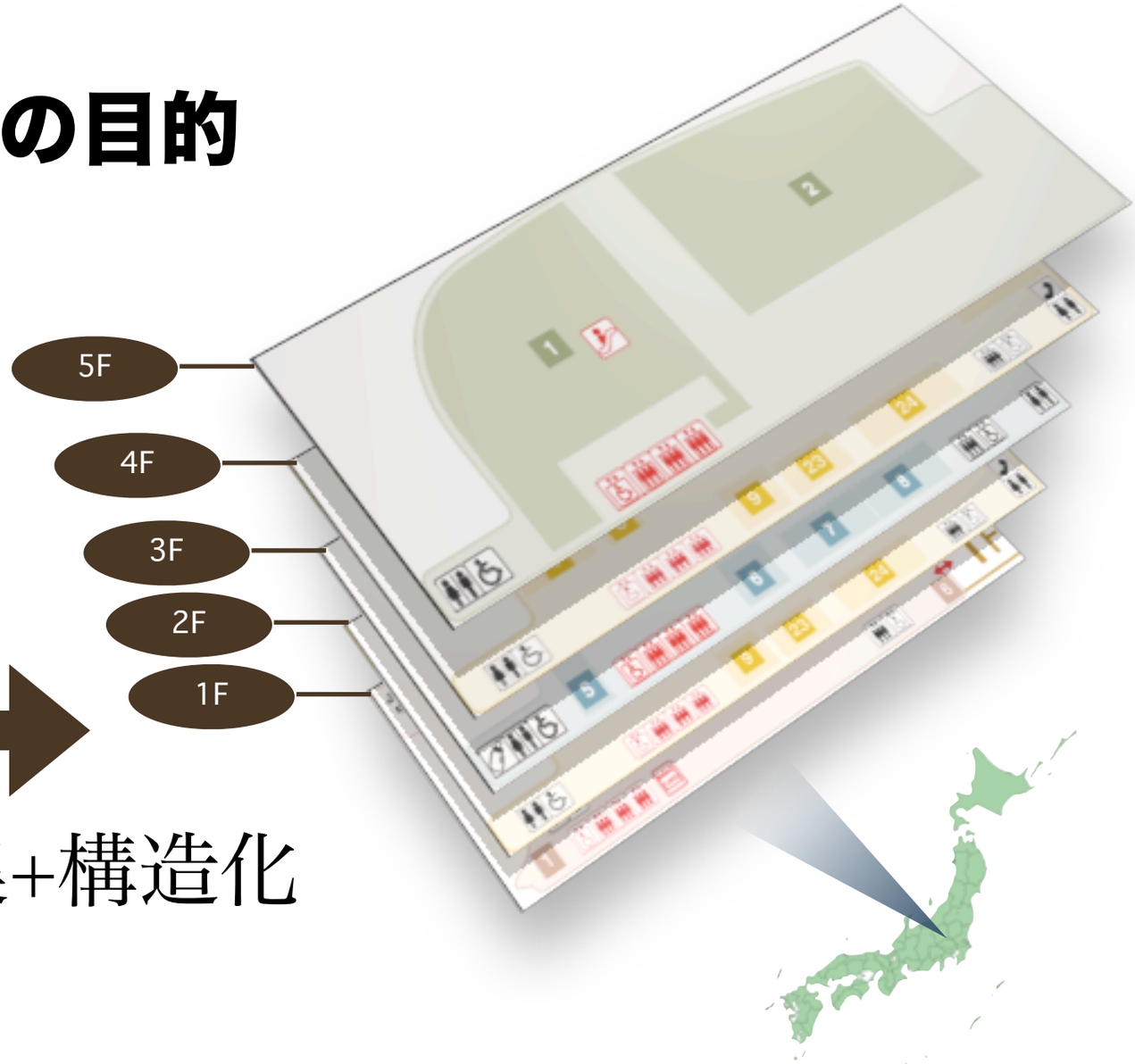
2013/03/07

マイクロジオデータ研究会

屋内地図情報収集のためのWebクローリング

フロアマップの収集システムの目的

収集+構造化

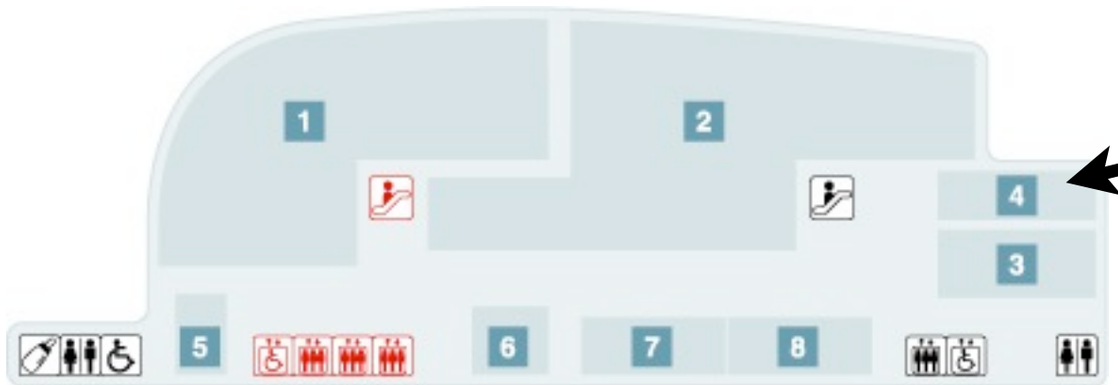


- 屋外の地図のように、航空写真やGPSログからの生成はできない
- 「フロアマップ」は公開されている
- ただし、Webページの「図」として貼り付けてあるだけ。空間的に整理されているわけではない。
- 収集して構造化できないか？

トップページ

フロアのインデックスページ

フロアマップ画像

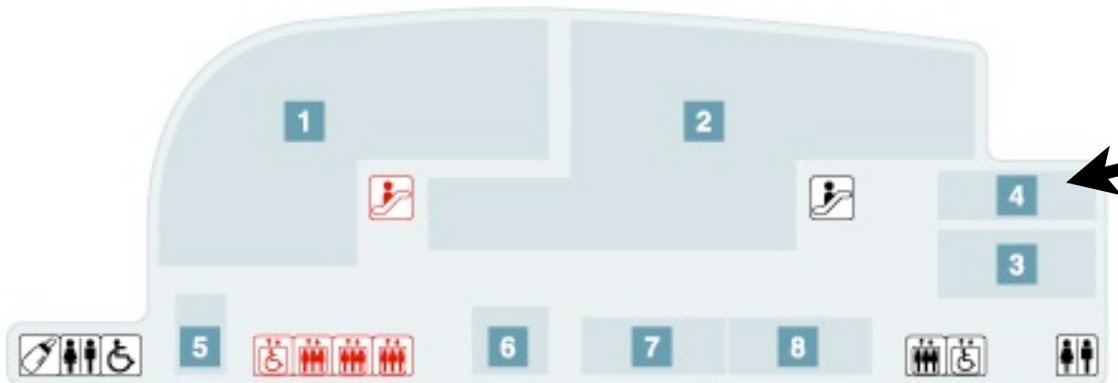


個別フロアのページ

トップページ

フロアのインデックスページ

フロアマップ画像



個別フロアのページ

フロアマップ収集の流れ

建物リスト作成 ← どの建物を対象とすべきか？

サイト検索 ← どのウェブサイトを対象とすべきか？

クローリング ← どのページを対象とすべきか？

フロアマップ画像抽出 ← どの画像がフロアマップか？

建物リスト作成

元データ

建物リスト

＋

別記（入居テナント一覧）

←… ゼンリンの住宅地図データから
対象となる建物をピックアップする

入居者（商業カテゴリ）ひとつにつき1点として数える

店舗らしいパターンの名前はもう1点追加する(計2点)

事務所らしい名前は加点しない

↓

抽出済み建物リスト

建物リスト例

#	aID	処理無し	事務所名フィルタ	店舗名加点	両方	抽出名
1	0042410900EA024500000039	新宿センタービル 227	ルミネエスト新宿店 222	新宿センタービル 275	■ ルミネエスト新宿店 251	ネエスト エスト
2	0042410900EB024500000750	ルミネエスト新宿店 222	マルイメン新宿玄海第二ビル 170	ルミネエスト新宿店 251	■ 新宿ＮＳビル 172	新宿ＮＳビル 宿ＮＳビル
3	0042410900EA024400000557	新宿パークタワー 217	新宿ＮＳビル 144	新宿パークタワー 240	■ マルイメン新宿玄海第二ビル 172	
4	0042410900EB0249000000FB	高田馬場ダイカンプラザ 215	新宿ミロード 125	高田馬場ダイカンプラザ 217	■ 新宿ミロード 140	ミロード 新宿ミロード
5	0042410900EA024600000974	新宿アイランドタワー 175	新宿パークタワー 117	新宿アイランドタワー 187	■ 新宿パークタワー 138	パークタワー ークタワー
6	0042410900EC024600000F77	マルイメン新宿玄海第二ビル 170	ルミネ１ 108	新宿ＮＳビル 185	■ ルミネ１ 136	新宿 宿
7	0042410900E9024300000A72	東京オペラシティビル東京オペラシティタワー 158	新宿三越アルコット 107	東京オペラシティビル東京オペラシティタワー 179	■ 新宿三越アルコット 120	新宿三越アルコット 宿三越アルコット
8	0042410900EA02450000002D	新宿ＮＳビル 154	新宿エルタワー 94	マルイメン新宿玄海第二ビル 172	■ 新宿エルタワー 120	支 エルタワー
9	0042410900EA024500000013	新宿住友ビルディング 152	高田馬場ダイカンプラザ 92	新宿住友ビルディング 170	■ 小田急百貨店本館 113	新宿 宿
10	0042410900EB0246000000A2	新宿ダイカンプラザＡ館 151	ルミネ２ 92	新宿ダイカンプラザＡ館 158	■ 新宿センタービル 111	ル 新宿センタービル
11	0042410900EA024600000410	新宿野村ビル 137	新宿センタービル 83	新宿野村ビル 156	■ ルミネ２ 110	新宿 宿
12	0042410900EA024500000016	新宿三井ビル 131	新宿野村ビル 82	新宿三井ビル 148	■ 新宿野村ビル 99	野村ビル 新宿野村ビル
13	0042410900EB0245000007AA	新宿ミロード 125	新宿プリンスホテル西武新宿ペペ 81	新宿ミロード 140	■ 高田馬場ダイカンプラザ 93	
14	0042410900EB0245000003B4	ルミネ１ 108	新宿ダイビル 76	ルミネ１ 136	■ 新宿プリンスホテル西武新宿ペペ 92	新宿ペペ 西武新宿ペペ
15	0042410900EB024500000736	新宿三越アルコット 107	新宿モノリス 75	新宿三越アルコット 120	■ 新宿ダイカンプラザＡ館 82	大ガード 新宿
16	0042410900EA02450000000F	新宿第一生命ビルディング 101	新宿ダイカンプラザＡ館 75	新宿エルタワー 120	■ 新宿モノリス 82	新宿モリノス 新宿モノリス
17	0042410900EB024500000217	レイフラット新宿 95	小田急百貨店本館 72	小田急百貨店本館 113	■ 東京オペラシティビル東京オペラシティタワー 81	オペラシティ ペラシティ
18	0042410900EA02460000072A	新宿エルタワー 94	新宿アイランドタワー 65	ルミネ２ 110	■ 新宿ダイビル 81	新宿アルタ 宿アルタ
19	0042410900EB024500000073	ルミネ２ 92	歌舞伎町ダイカンプラザ星座館 64	新宿第一生命ビルディング 109	■ 新宿住友ビルディング 79	住友ビル 新宿住友ビル
20	0042410900EA024700000711	ワコーレ新宿第一ビル 86	新宿住友ビルディング 63	レイフラット新宿 96	■ 飯田橋セントラルプラザ東京都飯田橋庁舎 76	飯田橋 田橋
21	0042410900EC024500000034	新宿Ｑフラットビル 86	東京オペラシティビル東京オペラシティタワー 80	新宿プリンスホテル西武新宿ペペ 92	■ 新宿アイランドタワー 73	アイランドタワー 新宿アイランド
22	0042410900EB02460000002C	新宿ダイカンプラザ７５６ 85	Ｌｅｅ３ 57	エステック情報ビル 88	■ 新宿三井ビル	新宿三井ビル

“外れ” が下位に押し下げられる

サイト検索

抽出済み建物リスト

建物名

建物サイトのURL

サイトを検索

Yahoo! API

（終了したため切り替え）

Google検索

APIではなくスクレイピングベース

※現在このスクレイピングは動かないかも……

抽出した建物名

もう一度検索

電話帳サイト、不動産情報のサイト等 = NGサイトが出てきた場合

住宅地図データ上の名称が適切でない可能性がある。

NGサイトリストに含まれている場合

建物名が一般的ではない可能性

→ 適切でない名称の例： 道玄坂共同ビル渋谷109 (渋谷109)

→ **テナント名に含まれる建物名を抽出**

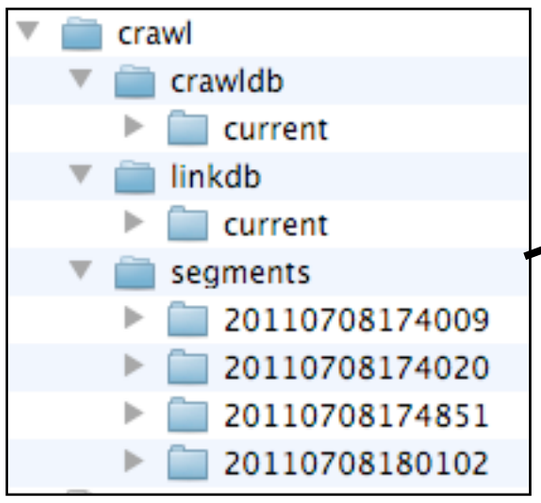
クローリング（STEP1：単純クローリング）

nutch を使用してクローリングを行う。

ここではフロアマップが掲載されているか判断せず、あらゆるページをクローリングする

・クローリングは（目的自体は）単純だが、実装は意外と大変。Web 上にはおかしな（汚い）ページがたくさんある。 それらにぶつかっても、クラッシュしないよう処理を続ける必要がある。

・今のところ nutch でクローリングして後からデータベースを抽出するという実装



nutch のクローリングデータベース（多数のファイル）を一つのDBファイルに変換し、自前のシステムではここからデータを取り込んで処理をする

クローリング（STEP2：ページのスコア計算）

ここからは、自前のシステムで処理を行う。

リンク追跡・評価

起点ページ（検索で見つかったページ）から順番にリンクを辿りながらリンクの評価をする。リンクに含まれるキーワードに応じて得点を付ける。

ページ単体評価

リンク同様、ページのタイトルやURLにもキーワードが含まれているかどうか調べ、得点を付ける。

→ページ単体の得点に、流入リンクの得点を合計してページの最終得点とする

→但し、ページの深さ（起点ページからのホップ数）により流入リンクを加算するか判断

リンク評価用スコア表

- keyword: フロア

  score:　3.0

- keyword: floor
  filter: downcase
  score:　3.0

- keyword: マップ

  score:　3.0

- keyword: map
  filter: downcase
  score:　3.0

- keyword: ガイド

  score:　1.0

- keyword: (b|B|地下)?[1-9]+[fF ｆ Ｆ 階]　← ここは正規表現によるパターンマッチ

  filter: regexp　（◯階 というパターンにマッチさせる　）
  score:　1.0

ページ評価用スコア表

- keyword: フロア

  score:　3.0

- keyword: floor
  filter: downcase
  score:　3.0

- keyword: マップ

  score:　3.0

- keyword: map
  filter: downcase
  score:　3.0

- keyword: ガイド

  score:　1.0

- keyword: (b|B|地下)?[1-9]+[fF ｆ Ｆ 階]

  filter: regexp
  score:　1.0

（現在のところ同じ）

ページスコア＋リンクスコア例

D	タイトル/URL	KW	Pr	Cmb
2	ヨドバシAkiba｜フロアガイド2-6F 19761B http://www.yodobashi-akiba.com/information/information_2_6.html	5.0 U 0.0 / T 5.0	40.0	**45.0**

7 Incoming Link(s)

[1] ヨドバシAkiba｜フロアガイド : K 7.0 - ヨドバシAkiba 6F : 1.0
[1] ヨドバシAkiba｜フロアガイド : K 7.0 - ヨドバシAkiba 5F : 1.0
[1] ヨドバシAkiba｜フロアガイド : K 7.0 - ヨドバシAkiba 4F : 1.0
[1] ヨドバシAkiba｜フロアガイド : K 7.0 - ヨドバシAkiba 3F : 1.0
[1] ヨドバシAkiba｜フロアガイド : K 7.0 - ヨドバシカメラ2F : 1.0
[2] ヨドバシAkiba｜フロアガイド2-6F : K 5.0 - ページTOP : 0.0
[2] ヨドバシAkiba｜フロアガイドB1-1F : K 5.0 - 2-6F : 1.0

2	ヨドバシAkiba｜フロアガイドB1-1F 22864B http://www.yodobashi-akiba.com/information/information_b1_f1.html	5.0 U 0.0 / T 5.0	16.0	**21.0**

4 Incoming Link(s)

1	ヨドバシAkiba｜フロアガイド 21820B http://www.yodobashi-akiba.com/information/information_floor.html	7.0 U 3.0 / T 4.0	7.0	**14.0**

67 Incoming Link(s)

1	ヨドバシAkiba｜アクセスガイド 17722B http://www.yodobashi-akiba.com/information/access_map.html	4.0 U 3.0 / T 1.0	6.0	**10.0**

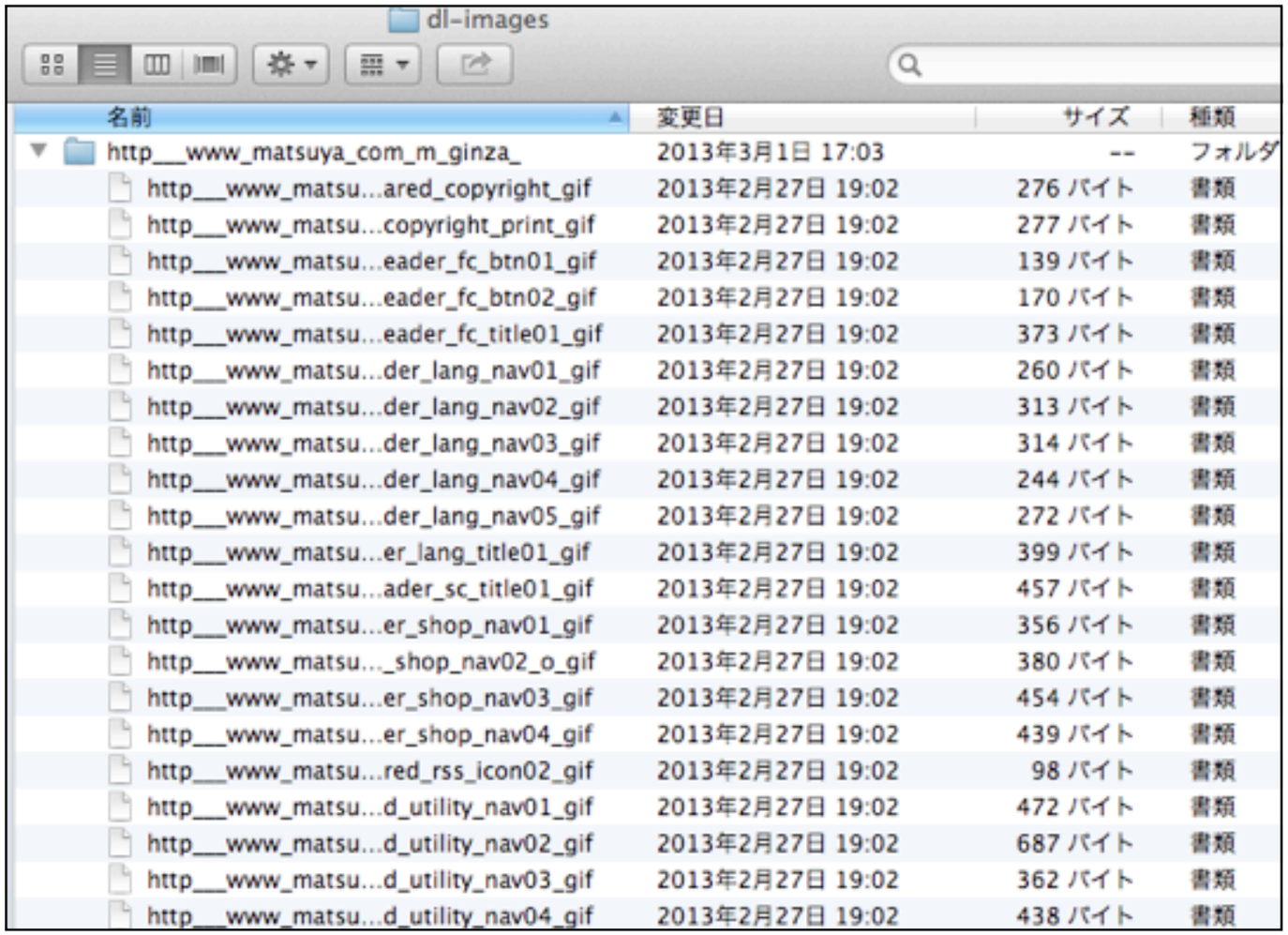
244 Incoming Link(s)

1	ヨドバシAkiba｜カフェ｜ビアードパパ	0.0	1.0	**1.0**

画像ダウンロード

候補のページに含まれる画像をダウンロードする

ただし、ページの得点が**0**であればダウンロードを行わない



ダウンロードしたディレクトリにJSONファイルを作り、スコア等を記録する

画像の抽出

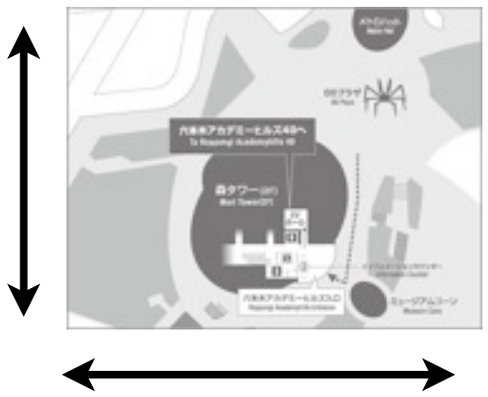
いくつかのアルゴリズムを組み合わせて画像の得点を計算する

画像寸法スコア	$Sg = \sqrt{w*h}$
関連テキスト キーワード	Sk
参照数	$Sr = 1/n$
SIFT特徴	Ss

$$Score = \{g\_normalize(Sg) + Sk + Ss\} Sr$$

* g_normalize(x) = x / Sgmax　ページ内で一番大きい画像のスコアで割る

画像評価-1　画像寸法



画像の幅、高さ

・画像の重要度に加味する

・極端に小さい画像を除外する（ボタンなど）

画像評価-2　関連テキスト キーワード

- 代替テキスト、タイトルテキスト、URLに含まれるファイル名の3つを対象にキーワードをマッチさせる
- キーワードはページ評価用のものと同じ

画像評価-3　参照数

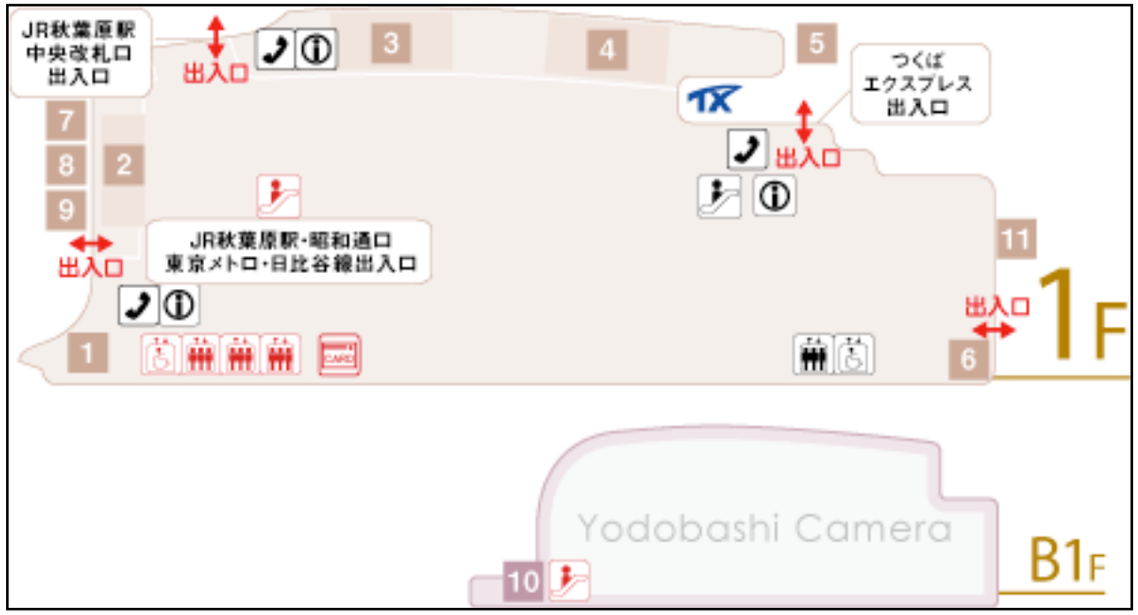
いろいろなページで参照されている画像は重要でない可能性が高い

画像評価-4　SIFT特徴

SIFT特徴により画像の分類を行う。分類結果のクラスごとに得点を割り当てる。

例

1st



寸法　　キーワード　SIFT　参照数

(1.0 + 10.0 + 2.0) / 1.0 = 13.0

2nd

(0.76 + 4.0 + 2.0) / 1.0 = 6.76

3rd

(0.43 + 8.0 + 0.0) / 3.0 = 2.81

フロア名抽出

何階のフロアマップであるか決定

- 画像の関連テキスト
- 画像のURL
- 画像を含むページのタイトル
- 画像を含むページのURL

(上にあるものが優先)

ページのスコア修正（次スライド）にも利用する

ページのスコア修正

画像の評価結果によりページのスコアを修正する

page_score’ = page_score * f

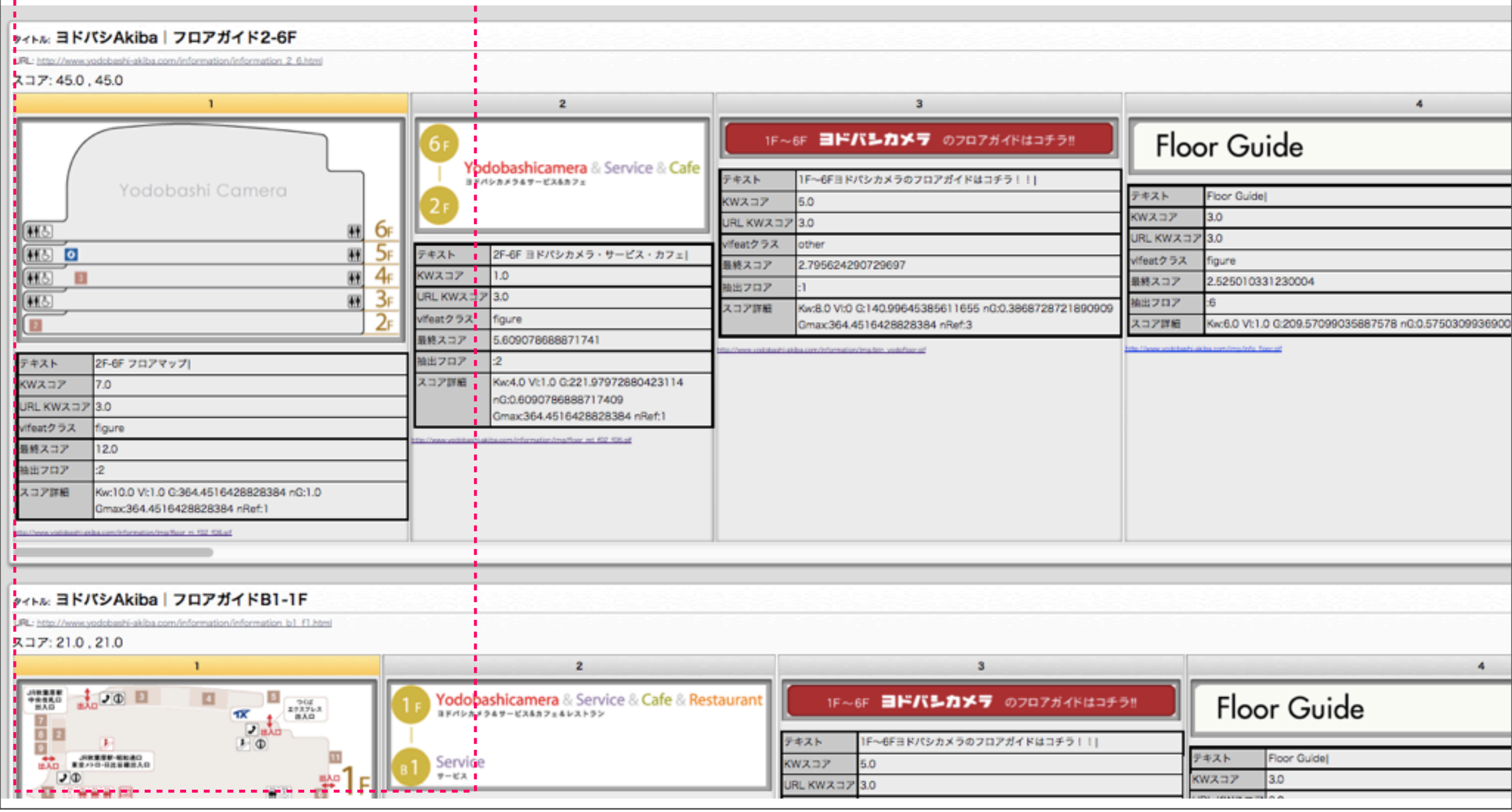
f = 1　その他

f = 0.5　画像のクラスが‘other’ またはフロア名が抽出されなかった場合

→ 例えば、ショップ一覧とフロアマップを掲載しているページが別々にある場合、フロアマップを掲載しているページが上位に来るようにする。 ショップ一覧の場合は店内の写真などが掲載されていると期待する。

収集例

各ページの最高スコア画像



主な課題

- フロア掲載ページとそれ以外の区別（スコアの足切り）
- 建物の通称（道玄坂共同ビル→渋谷109）の抽出
- 電話帳サイト、Wikipedia 等の排除
- PDF、Flash等（Flashは最近減ってきたので幸運）
- スコア、計算式の調整