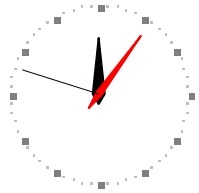
訓練集合の半自動生成に基づく専門検索エンジンの高速生成法

宮川 礼子　鈴木 悠生　鍋島 英知　岩沼 宏治

山梨大学

2006/8/5

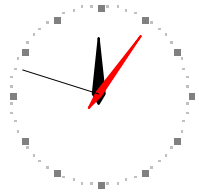
背景



Internet

大量の Web ページ

背景



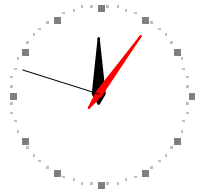
Internet

目的の情報
はごく一部

適切な検索結果を取得することは，いまだに困難

適切なキーワードからなる詳細な論理式が必要

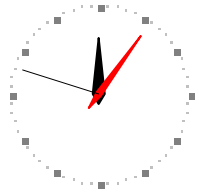
検索結果の品質向上の試み



- 汎用検索エンジンのパーソナライゼーション
  ⇒ 検索範囲を個人の趣味・嗜好に限定
- 専門検索エンジン
  ⇒ 検索範囲を特定のドメインに限定

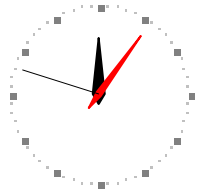
検索範囲を限定することで無関係なページを抑制する

専門検索エンジンの構築手法



- 手動構築
  - 管理者，ユーザが Web サイトを人手で登録
    ⇒ 面倒くさい ＆ 網羅性に欠ける
- 自動構築
  - 特定ドメインの Web ページをロボットにより収集
    ⇒ ロボット開発が大変 ＆ 設備費用がかかる
  - 汎用検索エンジンを用いて特定ドメインに関する Web ページを取得
    ⇒ 巨大なインデックスを作成･更新する必要がない
    ⇒ 最新の情報検索技術でランク付けされた検索結果を利用可能

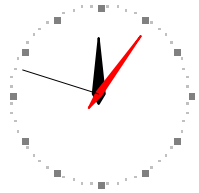
専門検索エンジンの構築手法



検索隠し味モデル
[小山ら, 04]

- 汎用検索エンジンを用いて特定ドメインに関する Web ページを取得
  - ⇒ 巨大なインデックスを作成・更新する必要がない
  - ⇒ 最新の情報検索技術でランク付けされた検索結果を利用可能

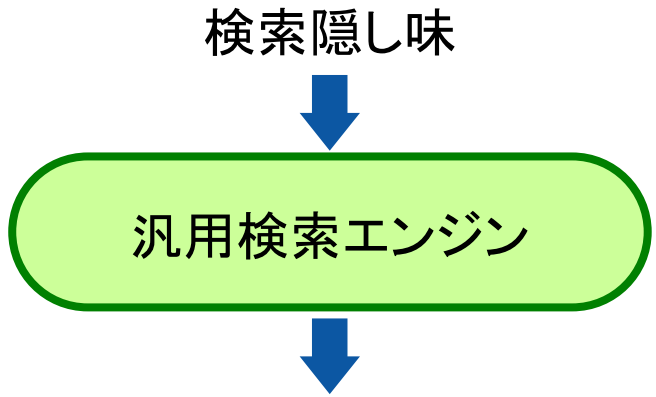
検索隠し味とは？ [小山ら, 04]



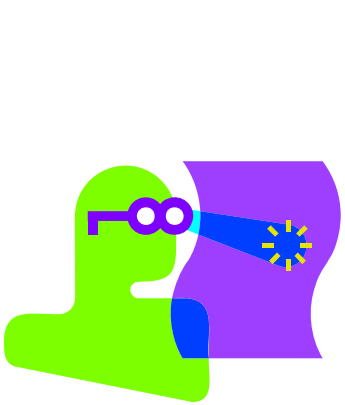
あるドメインに属する Web ページ群を特定するためのキーワードのブール式

例：料理レシピドメイン

（材料 ∧ ¬専門 ∧ ¬商品）∨ 大さじ



検索隠し味による専門検索エンジン



ユーザ質問
q

専門検索
エンジン

検索隠し味を付加
$q \wedge s$

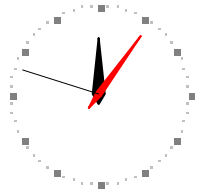
汎用検索
エンジン

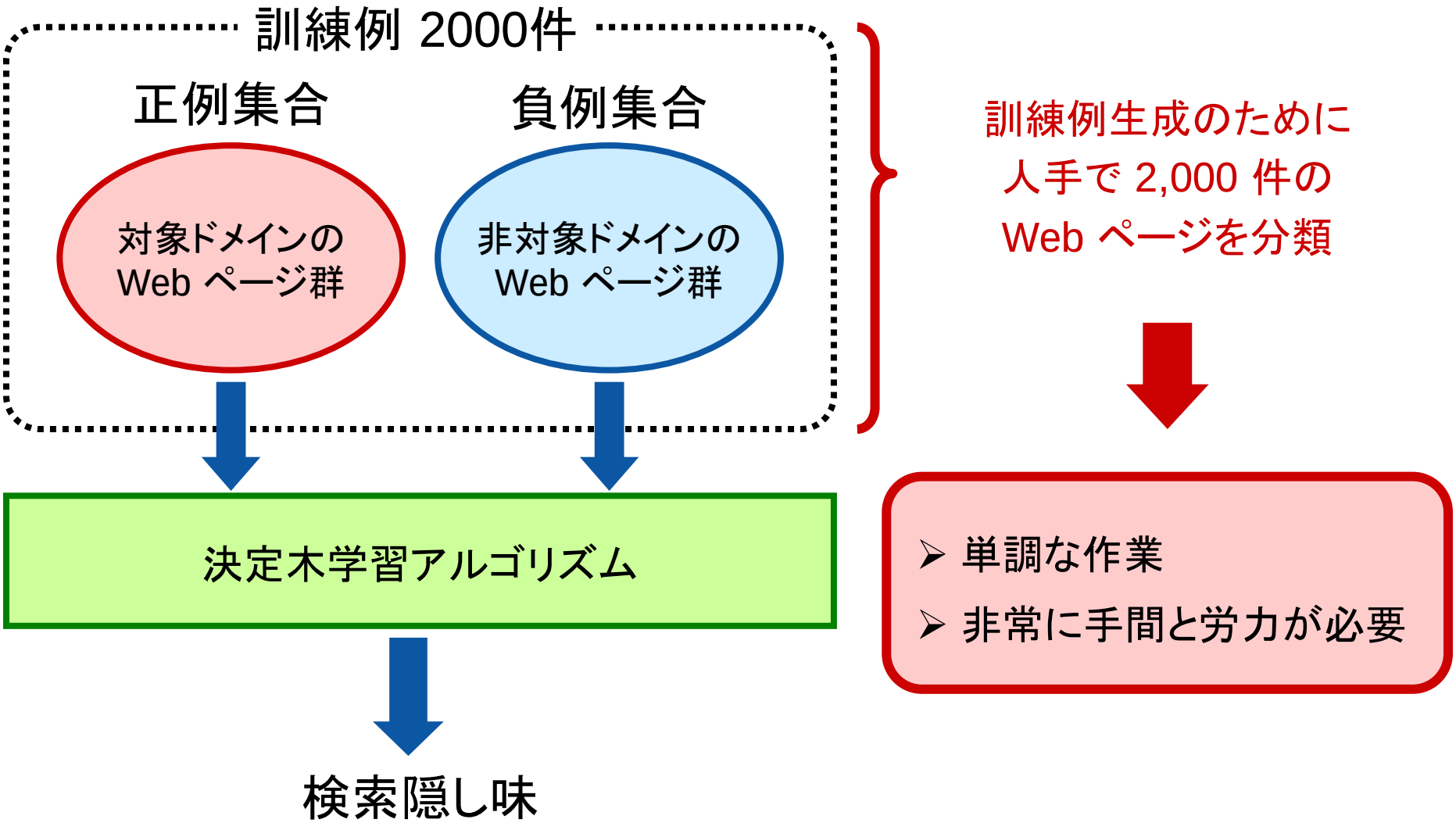
検索結果

対象ドメインの Web ページ群のうち，
ユーザ質問 q に適合する文書群

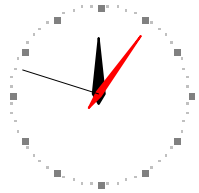
料理レシピを対象ドメインとしたとき，適合率 97% 以上，
再現率 86% 以上という非常に高い性能を示した [小山ら, 04]

検索隠し味モデルの問題点





研究目的

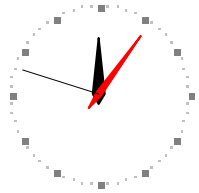


訓練集合の半自動生成

非常に少ない手間で専門検索エンジンを構築可能

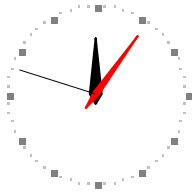
- 精錬による半自動生成法
  - Yahoo! などの Web ディレクトリから正例集合を精製
- 類似度に基づく半自動生成法
  - 極少数の正例をもとに正例集合を生成

発表の構成



- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

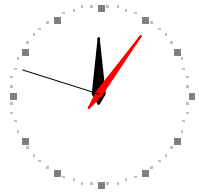
決定木の学習 [Quinlan, 86]



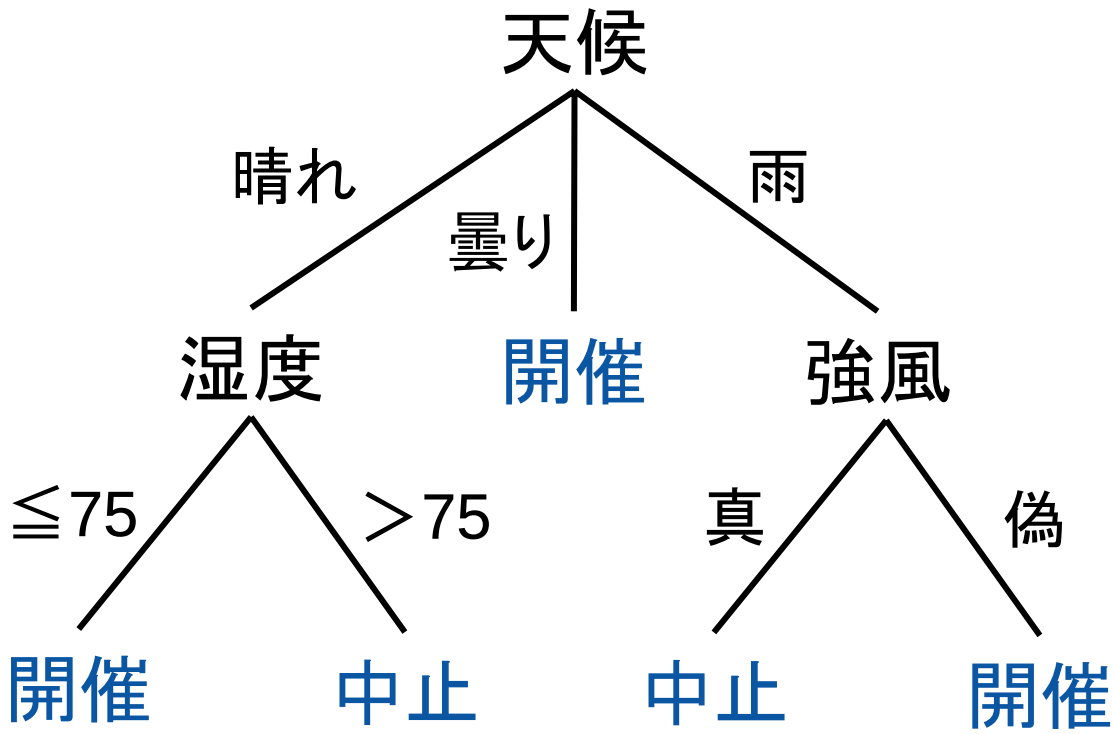
決定木：事例をクラスに分類する分類器

天候	温度	湿度	強風？	クラス
晴れ	75	70	真	開催
晴れ	80	90	真	中止
晴れ	85	85	偽	中止
晴れ	72	95	偽	中止
曇り	72	90	真	開催
曇り	83	78	偽	開催
曇り	64	65	真	開催
雨	71	80	真	中止
雨	65	70	偽	中止
雨	75	80	偽	開催
:	:	:	:	:

決定木の学習 [Quinlan, 86]

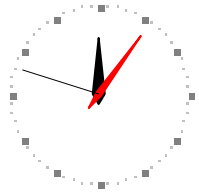


決定木：事例をクラスに分類する分類器



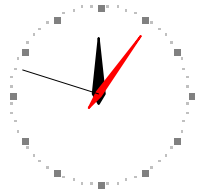
情報量利得比の観点から，適度に一般化された ＆ コンパクトな決定木を生成

発表の構成

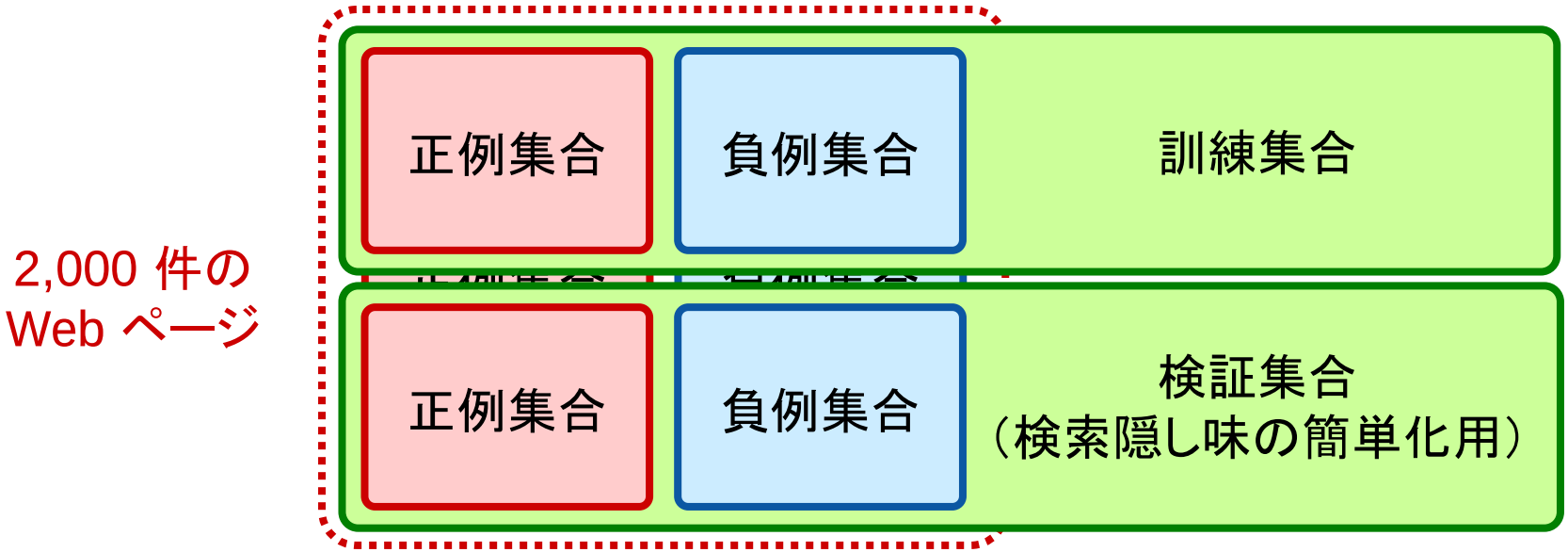


- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

検索隠し味の手動抽出 (1/2)
ドメインに属するページの収集



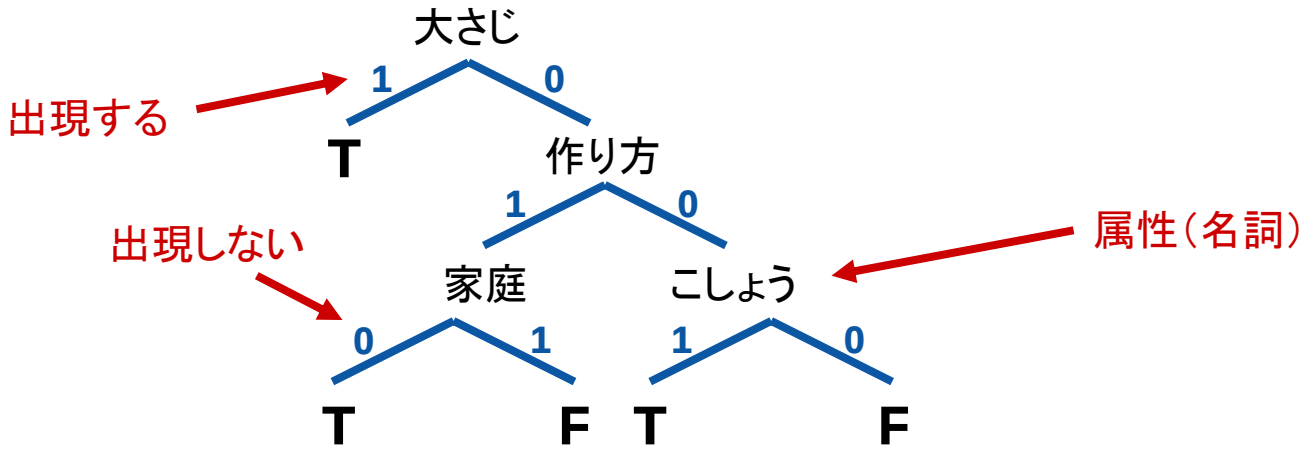
1. 将来ユーザが入力すると予想されるキーワードを 10 個選ぶ
   （牛肉，鶏肉，ピーマンなど）
2. 各キーワードを汎用検索エンジンに入力し，検索結果上位 200 件，合計 2,000 件の Web ページを収集
3. 人手により正例と負例とに分類し，訓練集合と検証集合を作成



検索隠し味の手動抽出 (2/2)

決定木の生成 & 検索隠し味の抽出

1. 各 Web ページから名詞を抽出し，ページの属性とする
2. 決定木学習アルゴリズムに訓練集合を与え，決定木を生成



3. 決定木を選言標準形（検索隠し味）に変換

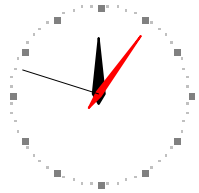
対象ドメイン　　非対象ドメイン

大さじ ∨（¬大さじ ∧ 作り方 ∧ ¬家庭）∨（¬大さじ ∧ ¬作り方 ∧ こしょう）

4. 検証集合を用いて検索隠し味を簡単化

根からたどるまでのパス

検索隠し味の性能



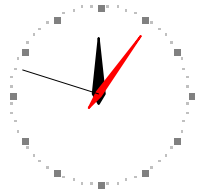
料理レシピドメインの検索隠し味

（材料 ∧ ¬専門 ∧ ¬商品）∨ 大さじ

	適合率		推定再現率
キーワード	キーワードのみ	検索隠し味を付加	検索隠し味を付加
豚肉	0.271	0.995	0.940
ほうれん草	0.205	0.976	0.870
エビ	0.063	0.986	0.976

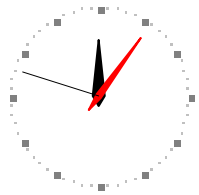
非常に高い適合率＆再現率を示した

発表の構成

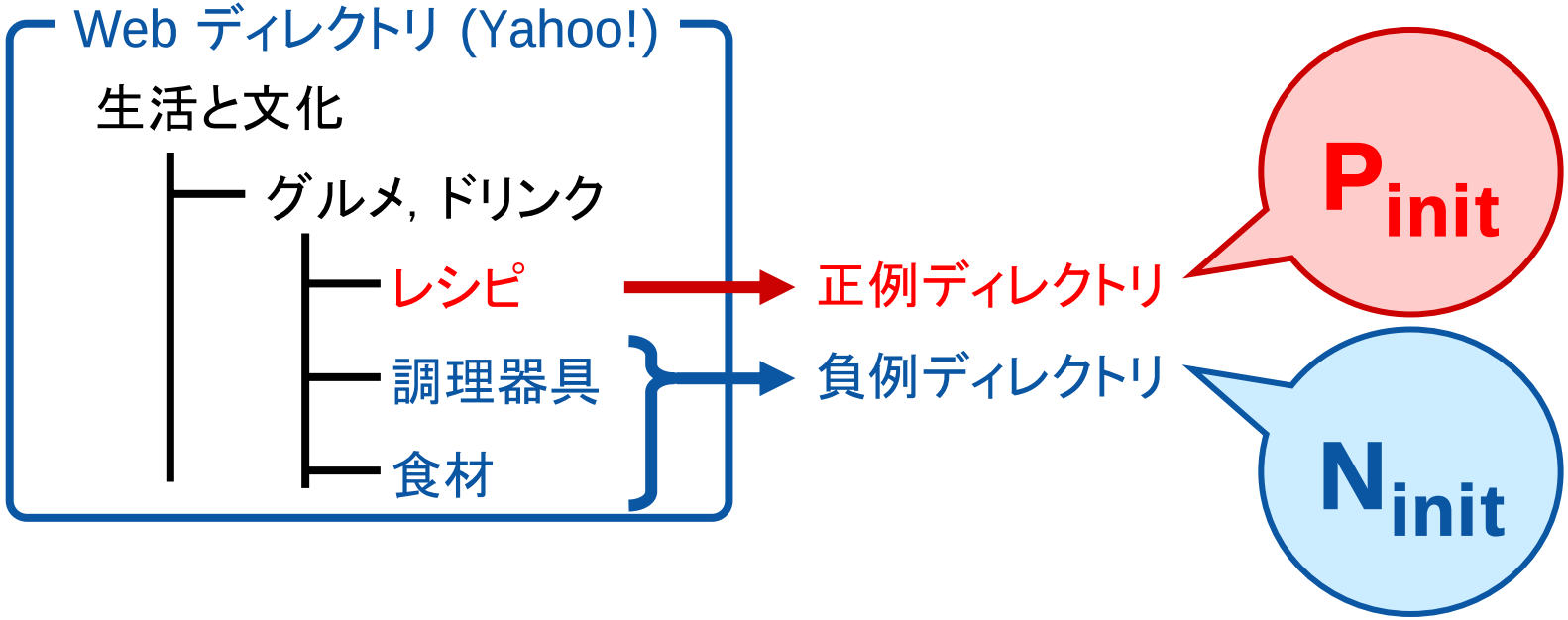


- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

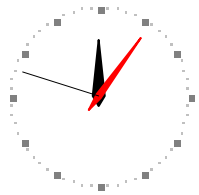
精錬による訓練集合の半自動生成 (1/4)
初期正例集合と初期負例集合の収集



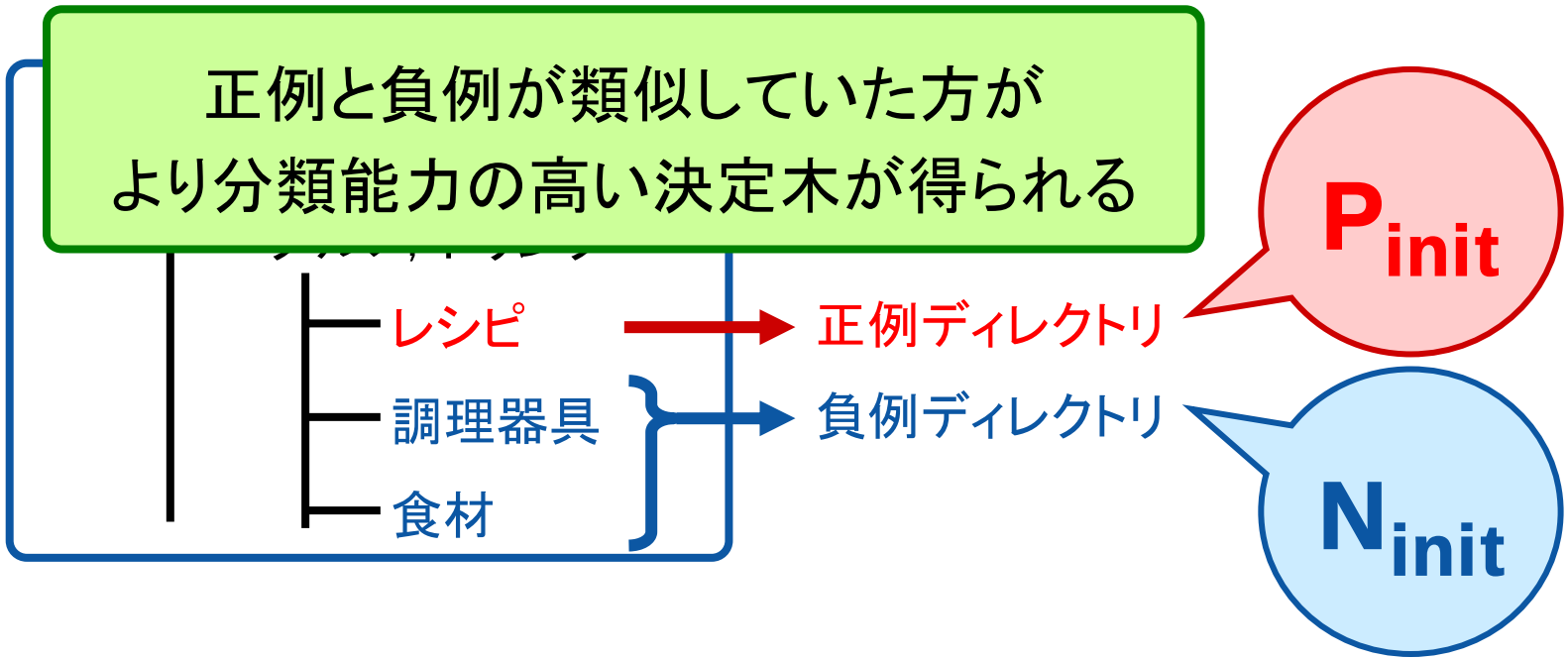
1. Web ディレクトリから対象ドメインに関するディレクトリ P_{dir} を選択
2. P_{dir} 中の Web サイト群を収集し，初期正例集合 P_{init} とする
3. P_{dir} の兄弟ディレクトリ中の Web ページ群を同様にして収集し，初期負例集合 N_{init} とする



精錬による訓練集合の半自動生成 (1/4)
初期正例集合と初期負例集合の収集

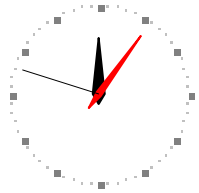


1. Web ディレクトリから対象ドメインに関するディレクトリ P_{dir} を選択
2. P_{dir} 中の Web サイト群を収集し，初期正例集合 P_{init} とする
3. P_{dir} の兄弟ディレクトリ中の Web ページ群を同様にして収集し，
   初期負例集合 N_{init} とする

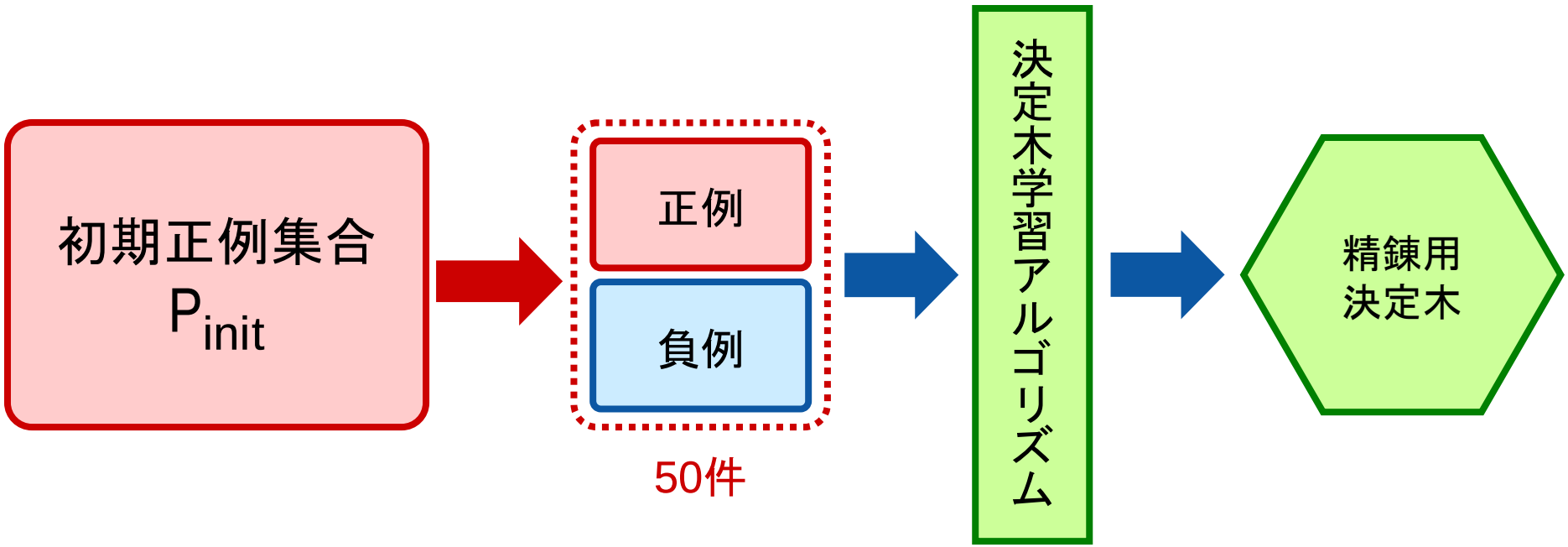


精錬による訓練集合の半自動生成 (2/4)

精錬用決定木の作成

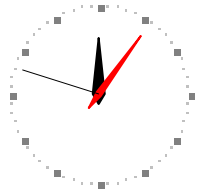


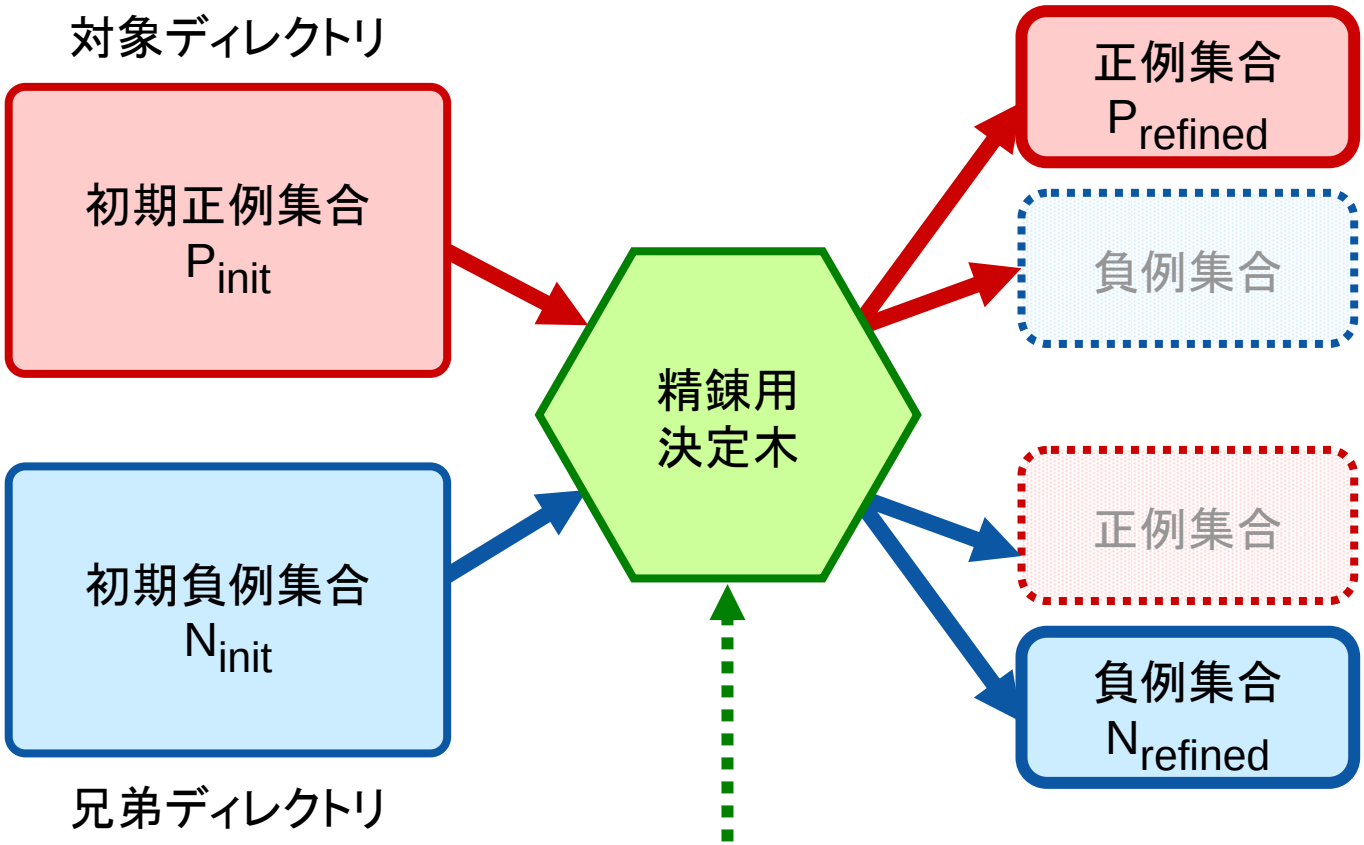
1. 初期正例集合 P_{init} から 50 件の Web ページをランダムに選択
2. 人手により，正例と負例とに分類
3. 決定木学習アルゴリズムにより，精錬用決定木を作成



精錬による訓練集合の半自動生成 (3/4)

初期訓練集合の精錬





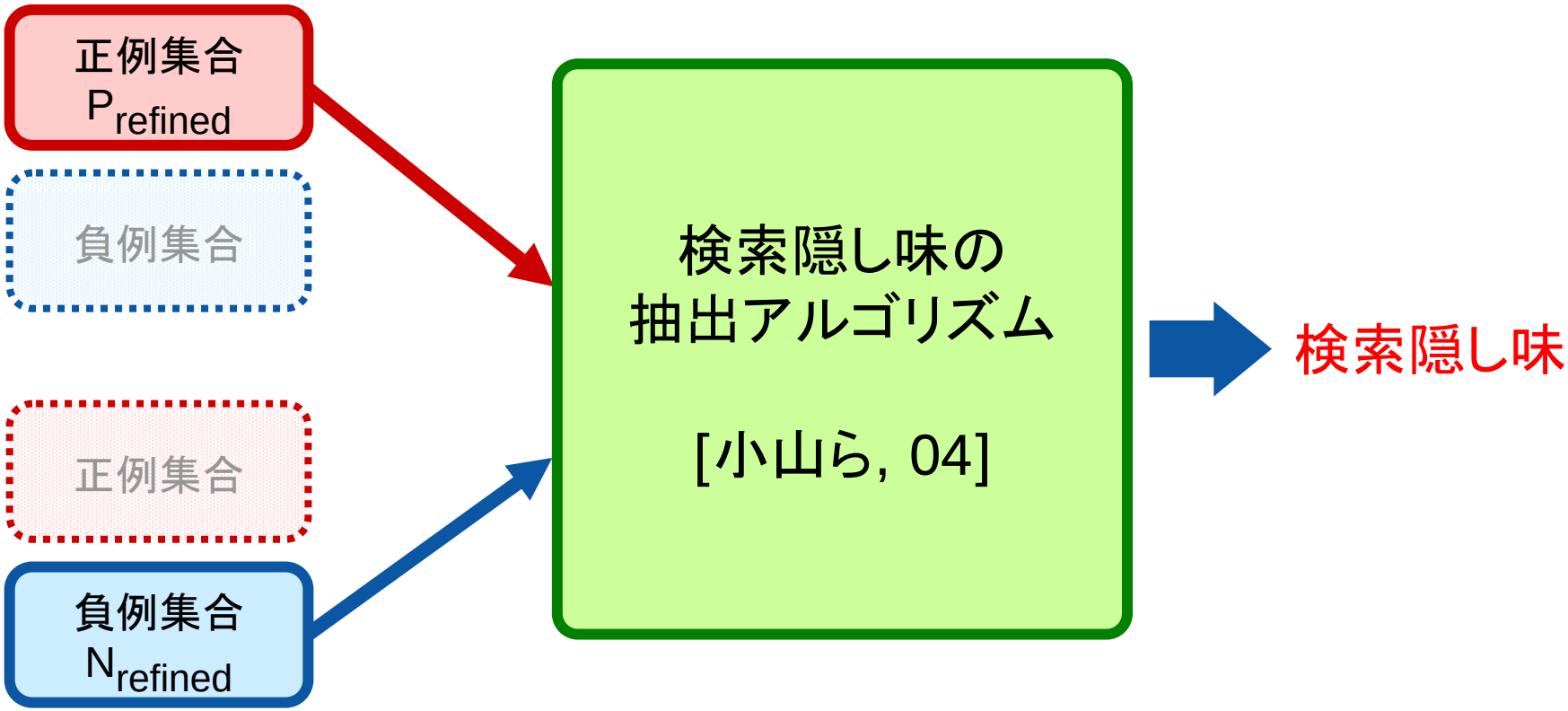
対象ドメインに
関係する
可能性が高い

ユーザの意図と
適合しない
Web ページ群

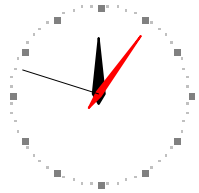
ディレクトリ管理者の意図と
ユーザの意図を識別するモデル

精錬による訓練集合の半自動生成 (4/4)

検索隠し味の抽出

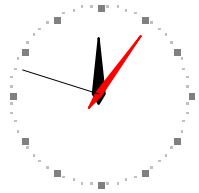


発表の構成



- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

評価実験1



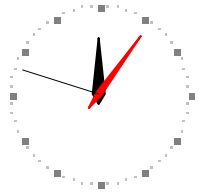
対象ドメイン

病気や怪我の詳細と治療法

比較実験

- 手動生成 ― 訓練集合を人手で作成（小山らの手法）
- 精錬半自動 ― 訓練集合の精錬あり
- 自動生成 ― 訓練集合の精錬なし

手動生成



10個のキーワード

肩，歯，鼻，膝，頭部，腰，胸部，目，心臓，血

Google に入力

検索結果上位 200 件

合計 2,000 件

人手で分類

正例集合

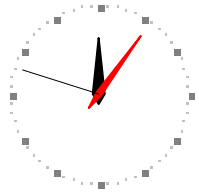
306 件

負例集合

1,694 件

検索隠し味

自動生成



“病気・症状” ディレクトリ

Yahoo!
ディレクトリ

初期正例集合
3,000 件

初期負例集合
7,000 件

兄弟ディレクトリ

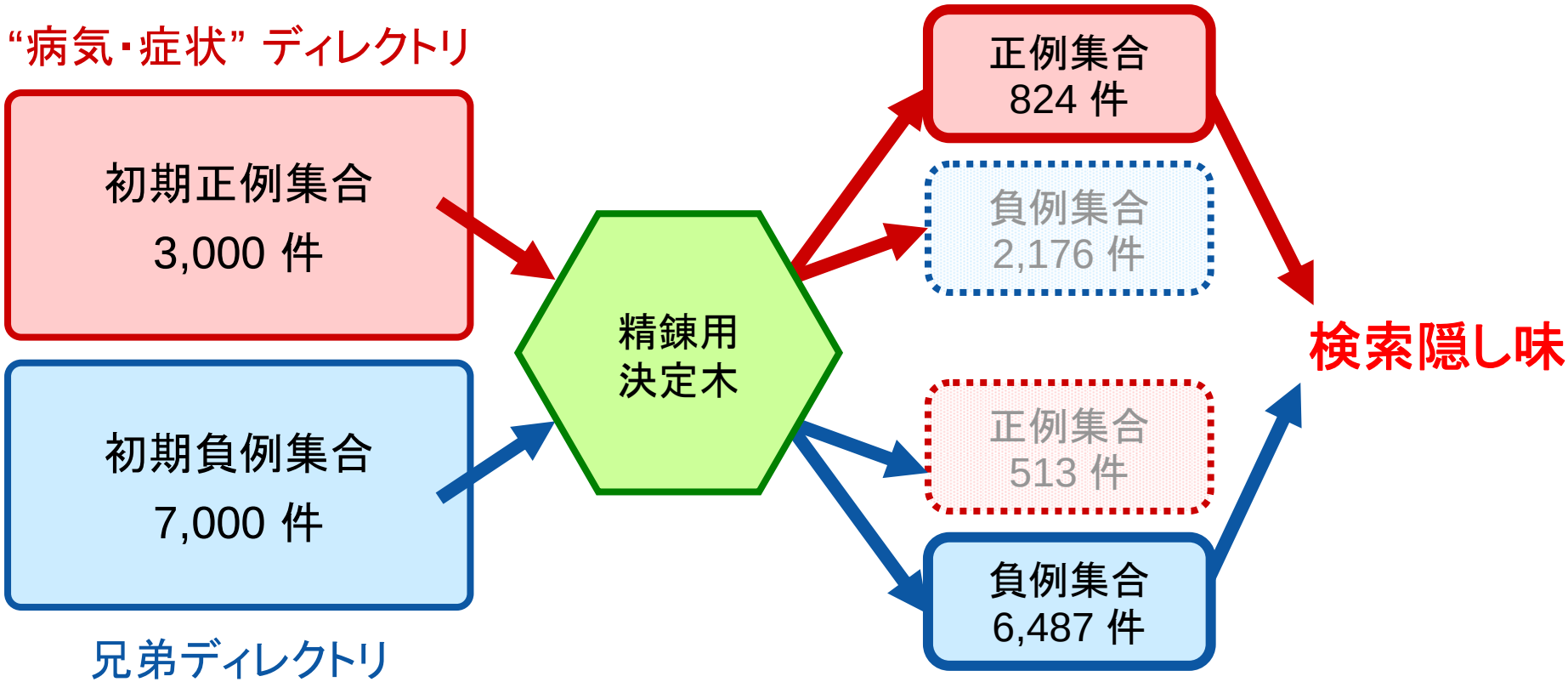
検索隠し味

Yahoo! ディレクトリ

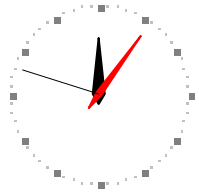
健康と医学

- 病気，症状
- 看護
- 職場

精錬半自動生成


“病気・症状” ディレクトリ
初期正例集合
3,000 件
初期負例集合
7,000 件
兄弟ディレクトリ
精錬用
決定木
正例集合
824 件
負例集合
2,176 件
正例集合
513 件
負例集合
6,487 件
検索隠し味

抽出された検索隠し味



	検索隠し味
手動生成	(症状 ∧ ¬注文 ∧ ¬チーム∧¬評価 ∧ ¬サイズ ∧ ¬デザイン ∧ ¬野 ∧ ¬町) ∨炎症 ∨ 頭痛
自動生成	¬パン ∧ ¬運営 ∧ ¬医薬品 ∧ ¬血清 ∧ ¬その他
精錬半自動	(症状∧¬紹介∧¬脳神経) ∨ 看病

Google にキーワードだけを入力した場合と，
検索隠し味を付加した場合の適合率と推定再現率を評価

検索結果上位 100 件の適合率

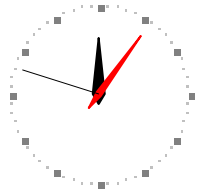
キーワード	キーワードのみ	手動生成	自動生成	精錬半自動
足首	0.08	0.69	0.10	0.60
頭	0.04	0.75	0.04	0.66
肺	0.22	0.73	0.26	0.70

推定再現率

キーワード	手動生成	自動生成	精錬半自動
足首	0.783	0.746	0.622
頭	0.794	0.786	0.703
肺	0.736	0.726	0.680

Web ページをほとんど絞り込めていないため

精錬による半自動生成法の安定性の評価

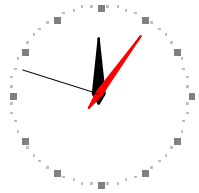


初期訓練例	検索隠し味
A (50件)	原因 ∨（状態 ∧ ¬闘病）
B (50件)	症状 ∧ ¬脳外科 ∧ ¬進歩 ∧ 児
C (50件)	（原因 ∧ ¬予定）∨ 食事
D (100件)	症状 ∧ ¬検索 ∧ ¬等
E (100件)	（症状 ∧ ¬ホームページ）∨ 発症

キーワード	A (50件)	B (50件)	C (50件)	D (100件)	E (100件)
足首	**0.47**	**0.63**	**0.48**	**0.63**	**0.61**
頭	**0.45**	**0.64**	**0.51**	**0.71**	**0.63**
肺	**0.68**	**0.60**	**0.69**	**0.69**	**0.63**

ばらつきが多い　　比較的安定

発表の構成

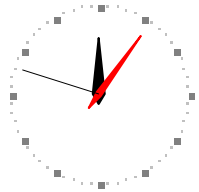


- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

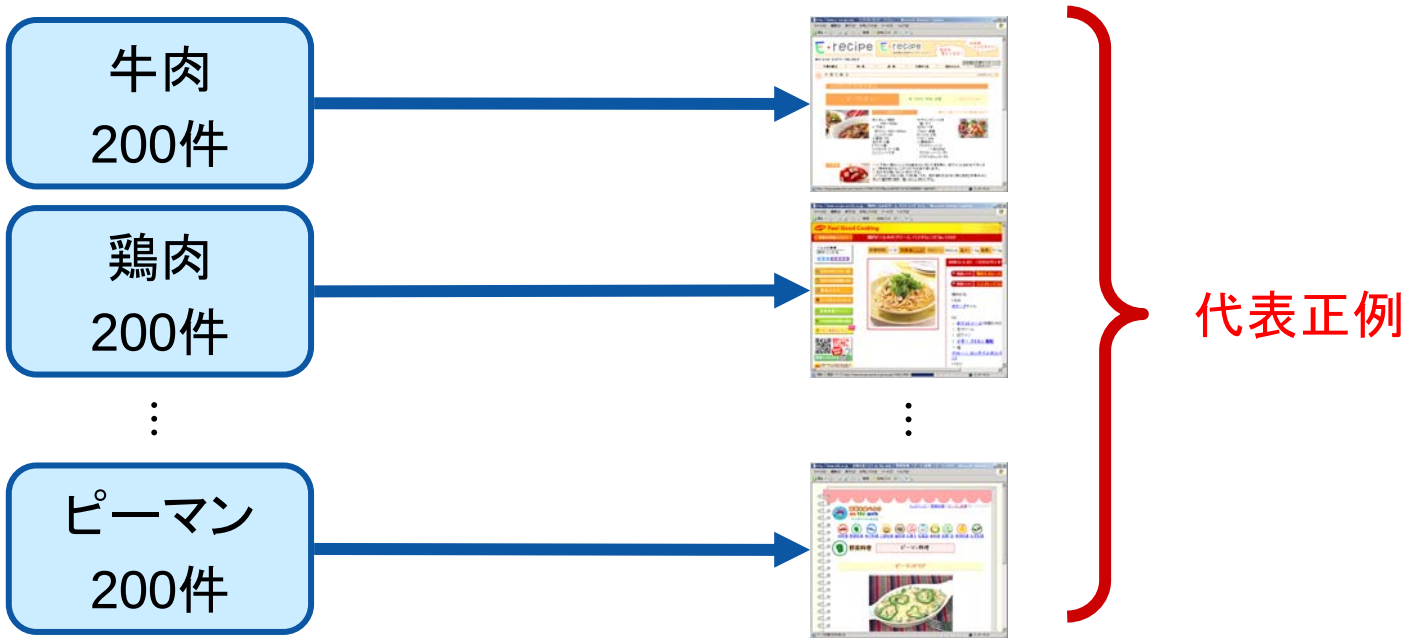
特徴

- Web ディレクトリにまったく依存しない
- 精錬による手法よりも人手がかからない

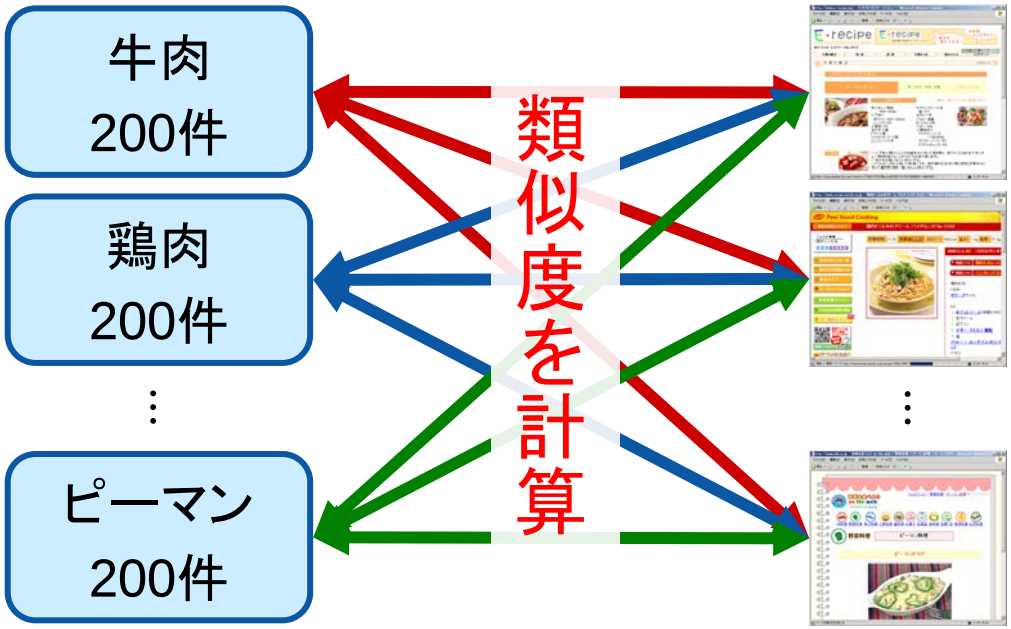
類似度に基づく訓練集合の半自動生成
代表正例の選択 (1/4)



1. 将来ユーザが入力すると予想されるキーワードを 10 個選ぶ
   （牛肉，鶏肉，ピーマンなど）
2. 各キーワードを汎用検索エンジンに入力し，検索結果上位 200 件，合計 2,000 件の Web ページを収集
3. 各キーワードに対し，適合するページを1つ選択（代表正例）

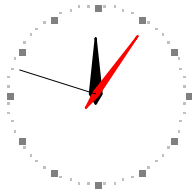


類似度に基づく訓練集合の半自動生成
類似度に基づく分類 (2/4)



代表正例との類似度を計算し，閾値以上ならば正例と分類

代表正例との類似度

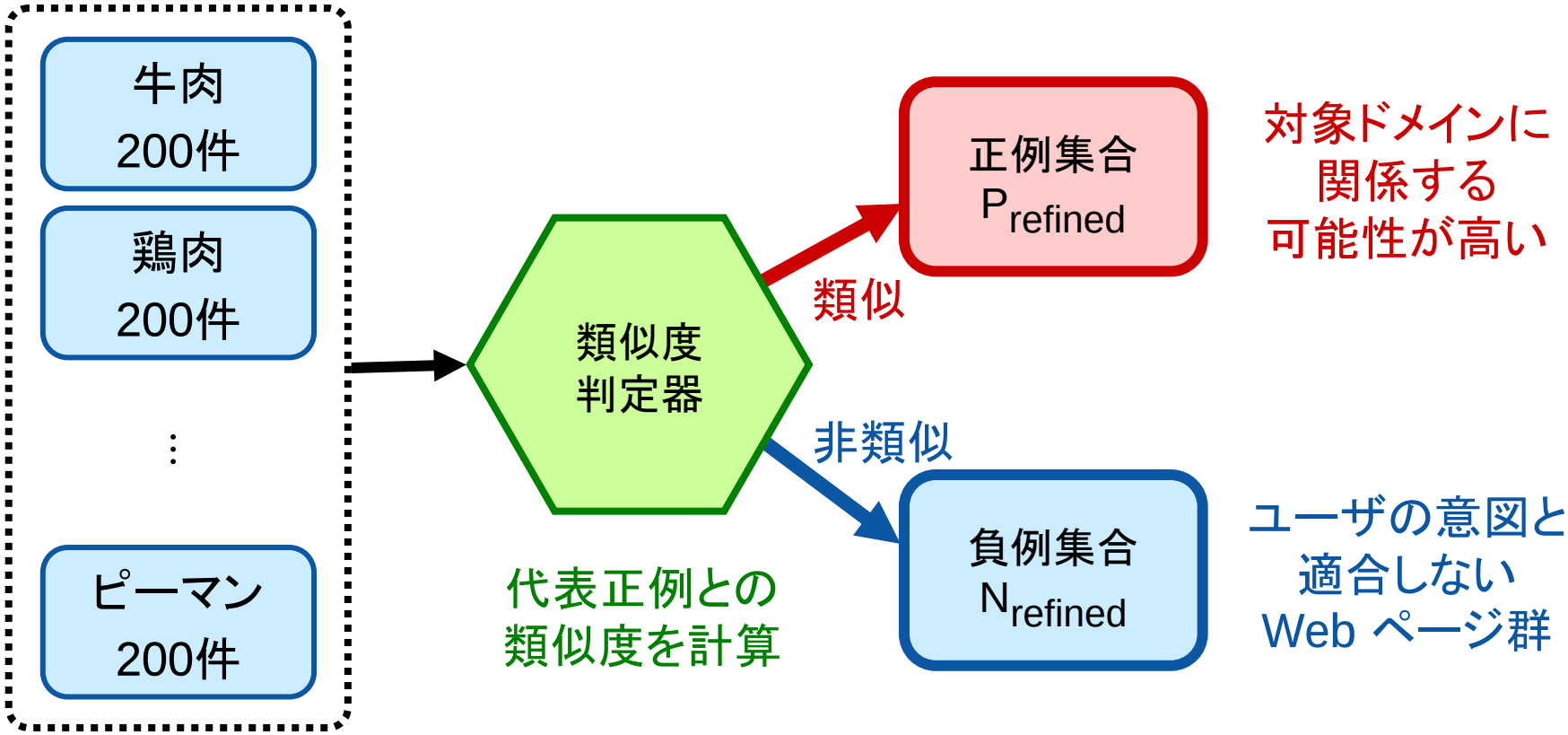


名詞の頻度ベクトルの余弦尺度により定義

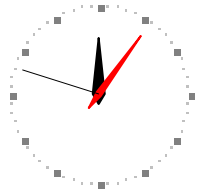
$$\cos(\boldsymbol{p},\boldsymbol{q}) = \frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{i=1}^{m} p_i^{\,2}}\sqrt{\sum_{i=1}^{m} q_i^{\,2}}}$$

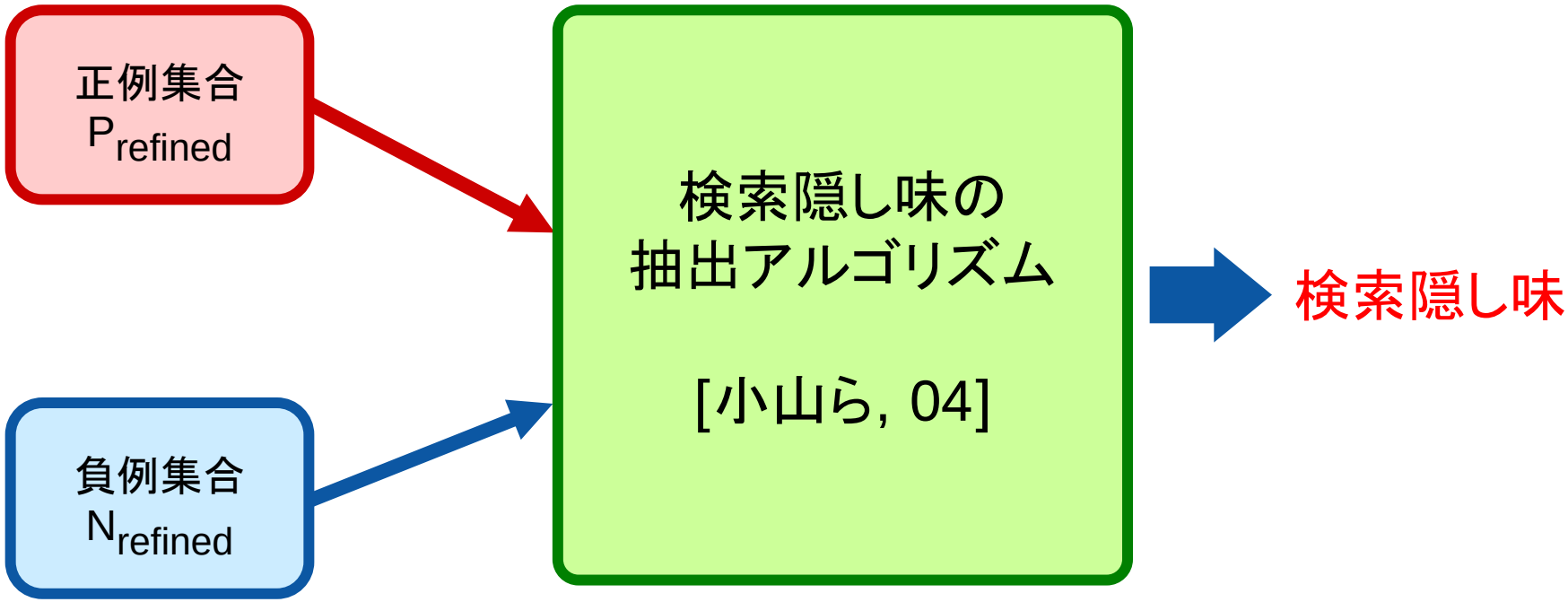
$\boldsymbol{p}, \boldsymbol{q}$	: Web ページ
p_i, q_i	: 名詞 i の出現頻度
M	: 名詞の種類数

類似度に基づく訓練集合の半自動生成
訓練集合の生成 (3/4)

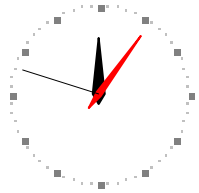


類似度に基づく訓練集合の半自動生成
検索隠し味の抽出 (4/4)



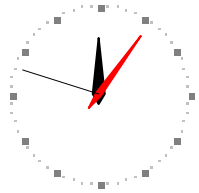


発表の構成



- 決定木の学習
- 検索隠し味の手動抽出法 [小山ら,04]
- 精錬による訓練集合の半自動生成
- 評価実験1
- 類似度に基づく訓練集合の半自動生成
- 評価実験2
- まとめ

評価実験2



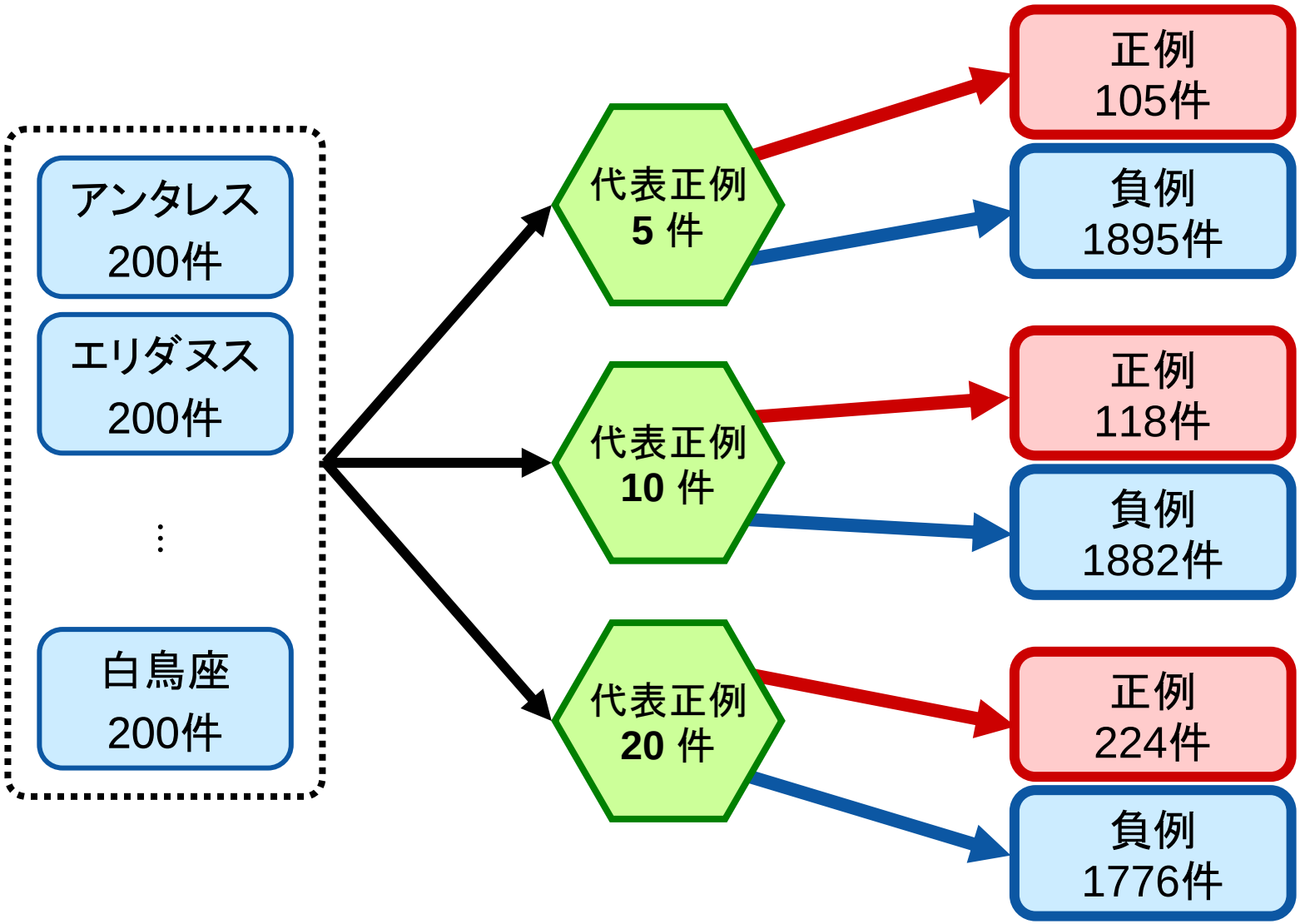
対象ドメイン

星の伝説，伝承

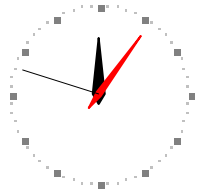
10個のキーワード

アンタレス，エリダヌス，アンドロメダ，ケンタウルス，
ペルセウス，牡牛座，大熊座，乙女座，琴座，白鳥座

代表正例数と分類精度


アンタレス
200件
エリダヌス
200件
白鳥座
200件
代表正例
5 件
代表正例
10 件
代表正例
20 件
正例
105件
負例
1895件
正例
118件
負例
1882件
正例
224件
負例
1776件

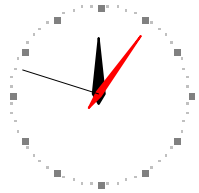
分類精度



		適合率
未分類		0.13
代表正例数5	正例	0.69
	負例	0.90
代表正例数10	正例	0.71
	負例	0.96
代表正例数20	正例	0.60
	負例	0.94

比較ページ数が増えるにつれ，分類能力が甘くなる

比較実験



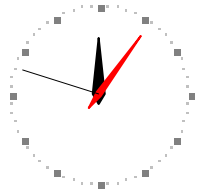
手動生成 ― 訓練集合を人手で作成
（小久保らの手法）

類似半自動5
類似半自動10 ― 類似度に基づく訓練集合の半自動生成
類似半自動20 （代表制例数5，10，20件）

精錬半自動 ― Web ディレクトリの精錬による半自動生成

Yahoo! には一致するディレクトリがないため，
関連の深い “生活と文化＞神話･民話と民俗学” を選択

抽出された検索隠し味



	検索隠し味
手動生成	神話 ∨ 退治 ∨ 大神
類似半自動5	アルゴル ∨ アルゴ
類似半自動10	エチオピア ∨ 物語 ∨ アルゴ
類似半自動20	(南中 ∧ ¬ヒドラ) ∨ (ペガサス ∧ ¬アルゴル)
精錬半自動	ゼウス ∨ (神話 ∧ ¬データ ∧ ¬神前 ∧ ¬国際)

Google にキーワードだけを入力した場合と，
検索隠し味を付加した場合の適合率と推定再現率を評価

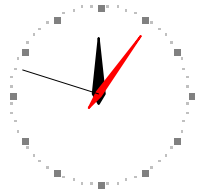
検索結果上位 100 件の適合率

キーワード	キーワードのみ	手動生成	類似半自動5	類似半自動10	類似半自動20	精錬半自動
オリオン	0.05	0.90	0.38	0.60	0.32	0.69
カシオペア	0.03	0.72	0.34	0.59	0.15	0.58
射手座	0.04	0.53	0.19	0.30	0.12	0.36

推定再現率

キーワード	手動生成	類似半自動5	類似半自動10	類似半自動20	精錬半自動
オリオン	0.94	0.03	0.97	0.30	0.16
カシオペア	0.97	0.01	0.76	0.00	0.02
射手座	0.48	0.00	0.68	0.00	0.09

まとめ



- Web ディレクトリを用いた精錬による訓練集合の半自動生成法
- ごく少数の正例との類似度に基づく訓練集合の半自動生成法

↓

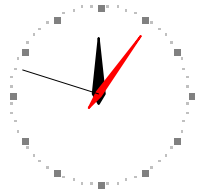
ユーザの所望するドメインに関する
検索エンジンを高速に構築可能

- 実験結果は，検索結果の質を大幅に改善できることを示している

↓

実用化に向けた大きな可能性を持っている

今後の課題



- 本手法に基づく専門検索エンジン構築サービスの実装
- 検索結果の品質のさらなる向上
- 他ドメインに適用し有用性を実証