Context-Awareな Web サービスをクラタリングするための単語類似度計算

Calculating Word Similarity for Context Aware Web Service Clustering

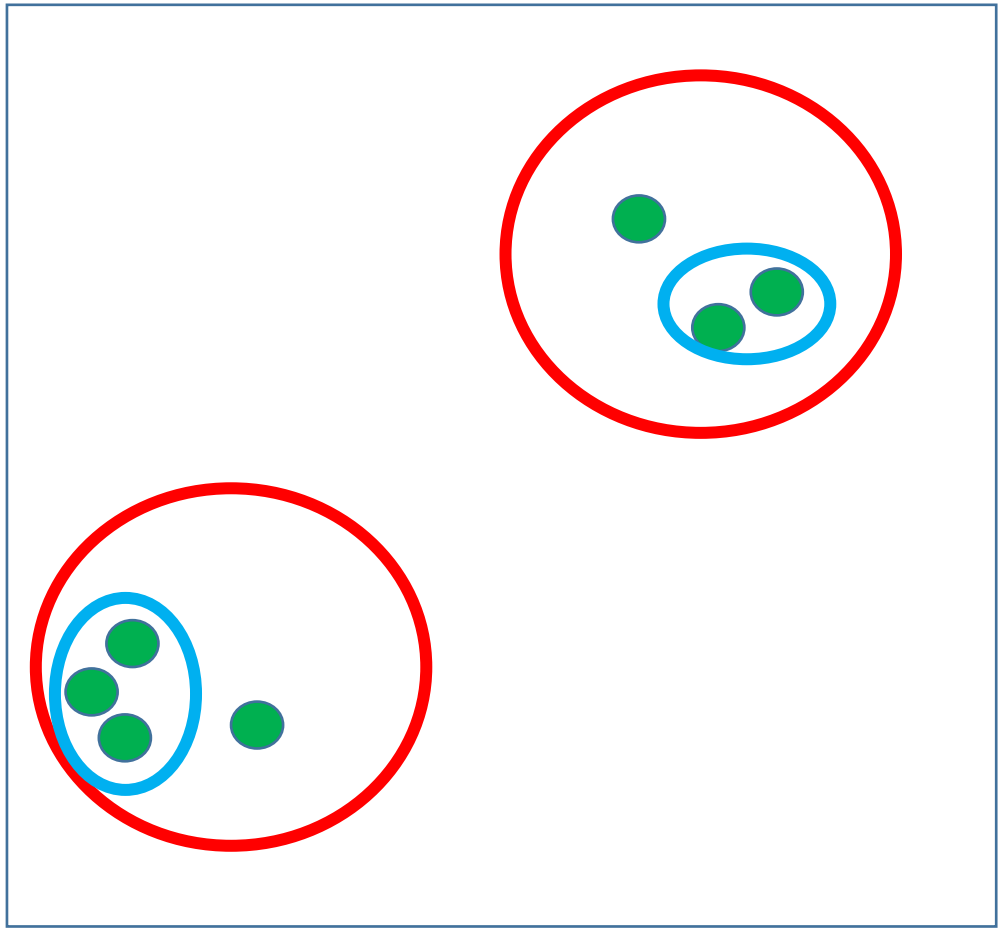
m5161136 Hiroki Ohashi

Web Service Discovery

- 様々なWebサービス増加し、混在する中で必要なサービスを見つけ出すのは難しい。
- これを解決するためにWebサービスをクラスタリングすることが非常に有効である。

Clustering

- クラスタリング
  - トップダウン
    - 全体の分割から順番に
  - ボトムアップ
    - 局所的なグルーピングから順番に



Web Service Similarity

- どのようにWebサービス間の距離（類似度）を調べてクラスタリングをしていくのか？
- WSDL等のWebサービス記述文書から各サービスの特徴を抽出し、それを利用して類似度計算
- 例）2つの電車情報サービス

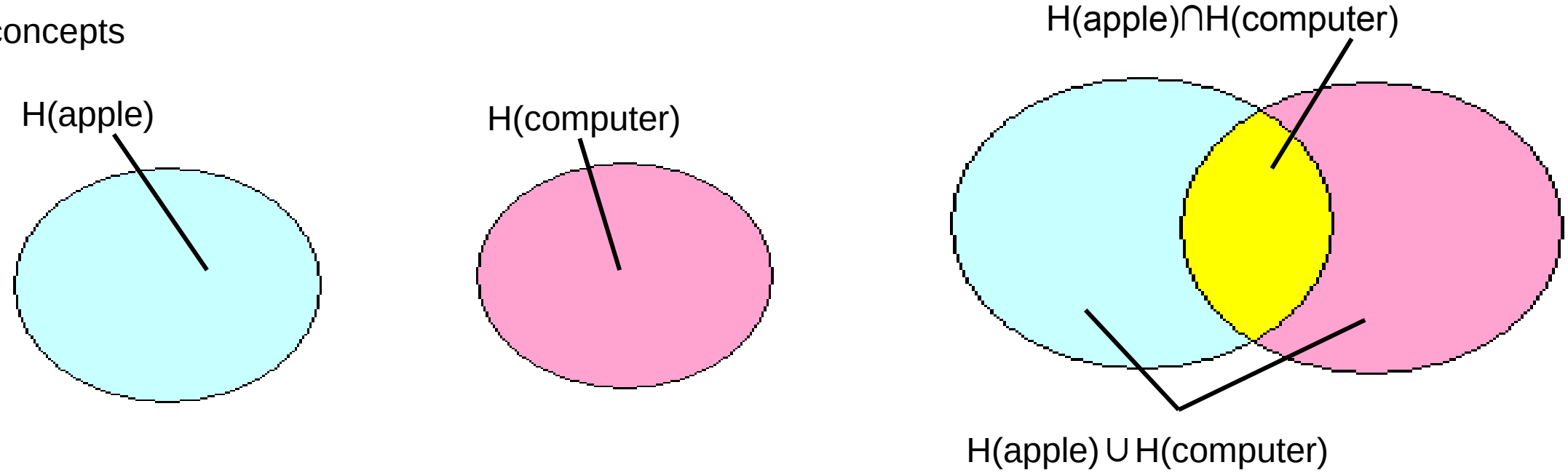
Service Name	Input1	Input2	Output
getTrainInformation	Station		Date
getExpressSchedule	DepartureCity	ArrivalCity	Date

Term Similarity

- ここでは単語間の距離を測る手法に着目する
- これまでに辞書やシソーラス（類語集）,サーチエンジンの結果などの様々な言語資源とコサイン類似度、TF-IDF等の計算アルゴリズムを組み合わせた多くの手法が研究されてきた.
- 特にサーチエンジンを利用した類似度計算は辞書を使った計算では分からない新たな関係性も高い類似度で表すことができる.

Hit count by Search Engine

・concepts



Web Similarity

- ヒットカウントを利用したSimilarity計算

$$\mathrm{WebJaccard}(P,Q) = \begin{cases} 0 & \text{if } H(P\cap Q) \leq c \\ \frac{H(P\cap Q)}{H(P)+H(Q)-H(P\cap Q)} & \text{otherwise} \end{cases}.$$

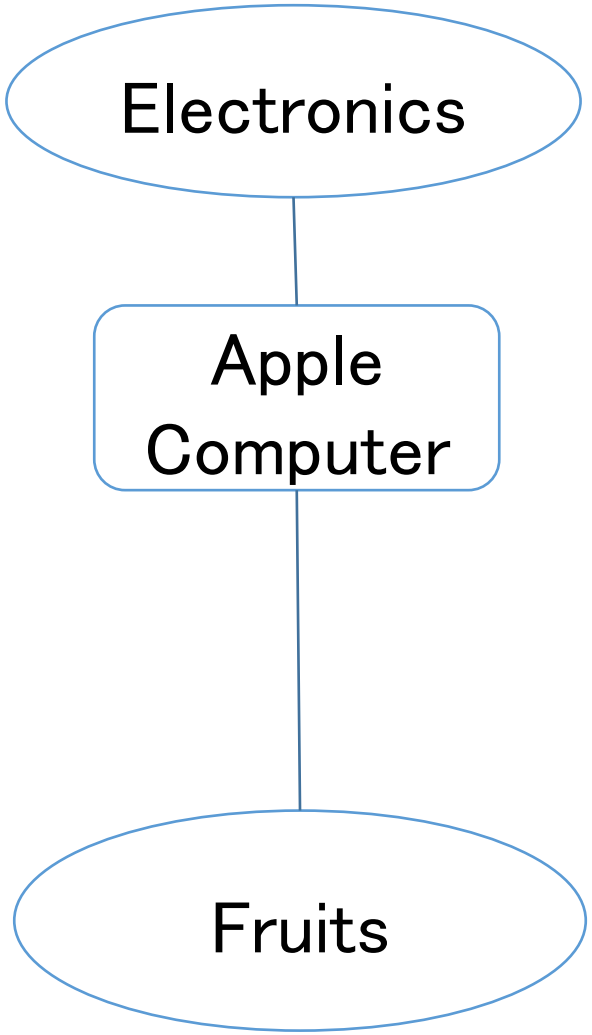
$$\mathrm{WebDice}(P,Q) = \begin{cases} 0 & \text{if } H(P\cap Q) \leq c \\ \frac{2H(P\cap Q)}{H(P)+H(Q)} & \text{otherwise} \end{cases}.$$

$$\mathrm{WebOverlap}(P,Q) = \begin{cases} 0 & \text{if } H(P\cap Q) \leq c \\ \frac{H(P\cap Q)}{\min(H(P),H(Q))} & \text{otherwise} \end{cases}.$$

$$\mathrm{WebPMI}(P,Q) = \begin{cases} 0 & \text{if } H(P\cap Q) \leq c \\ \log_2\left(\frac{\frac{H(P\cap Q)}{N}}{\frac{H(P)}{N}\frac{H(Q)}{N}}\right) & \text{otherwise} \end{cases}.$$

Term Similarity for domain

- 例えばappleとcomputerという単語ペアでは、一般的に類似度が高くなる.
  しかし、どの分野(ドメイン)においてどのぐらい近いのかまでは分からない.
- ここでは単語ペアをあるドメインで考慮した類似度を出すことを目標.
- appleとcomputerの関係性というのはelectronicsという観点からは高いものと考えられるが、fruitにおいては低くなる.



Motivation

- 各Web サービスにおいても同様に近いドメインと遠いドメインが存在する.
- 例
  - getTransportationLocationでは
  Trip > Geography > Food
  - getBreadPriceではFood>Trip = Geography
- ドメインについて考慮された類似度計算をサービス比較に適応することによってサービスのクラスタリングや検索の精度を高められると思われる.

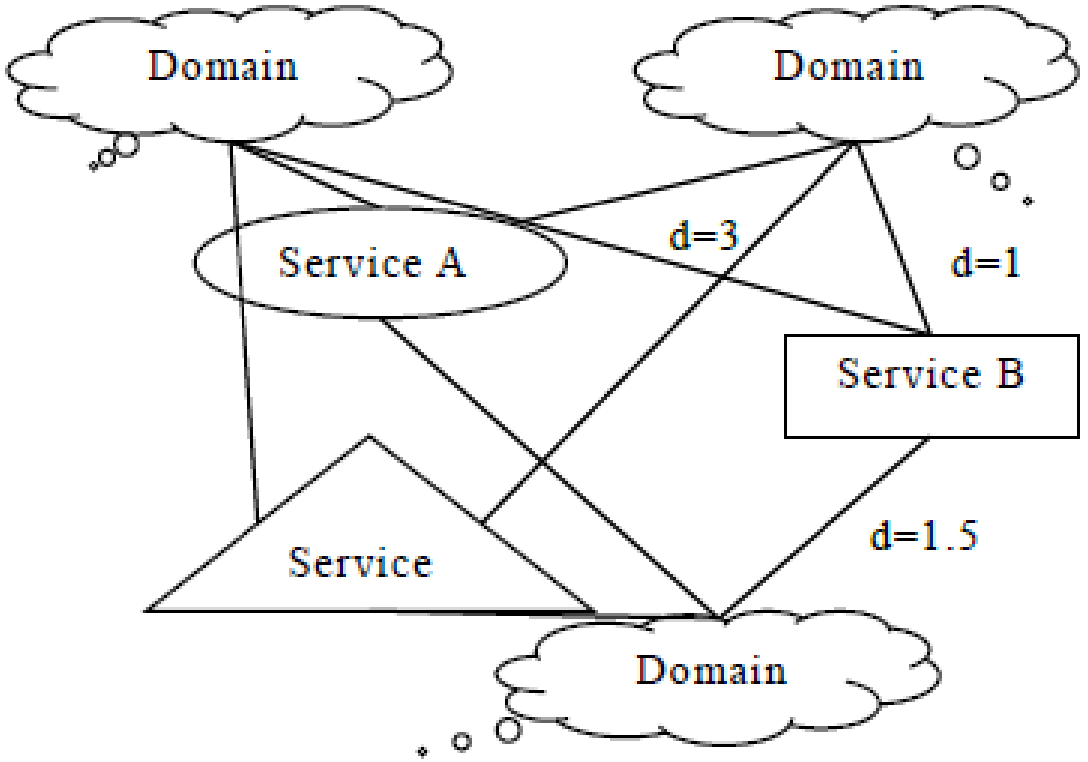


図 1　ドメインとサービスの関係

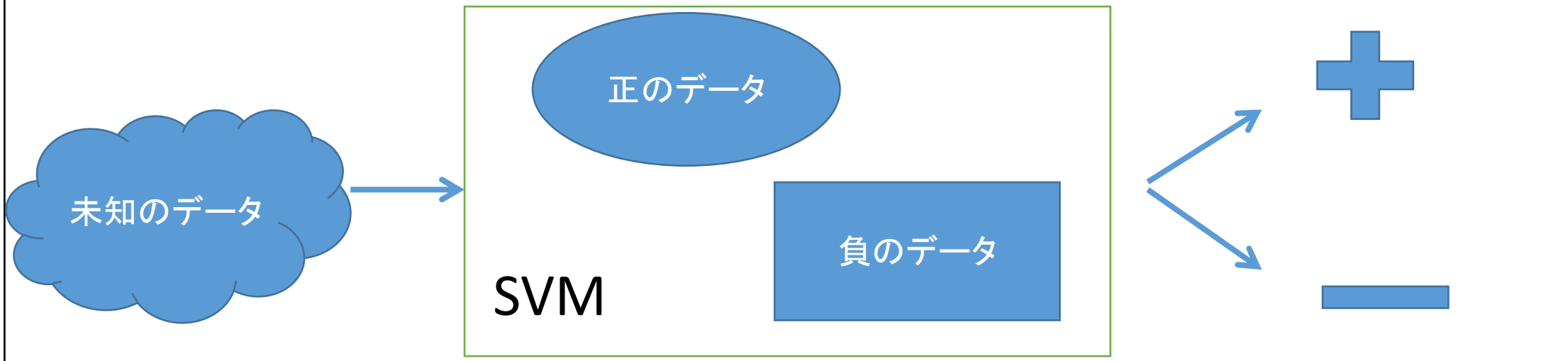
Approach

- 言語資源としてWeb検索の結果を利用(スニペット).
- ドメイン毎にそのドメインに属する2つの単語間で特徴ベクトルを作り、Support Vector Machine(SVM)で学習、分類させる.

- Step1: ドメイン毎の特徴ベクトルの決定
- Step2: 学習データの準備
- Step3: SVMによる分類

Support Vector Machine

- 機械学習の一つで学習データをもとにモデルを作成、未知のデータを高い精度で分類することができる.
- ここでは実際にドメインに属している単語ペアを正の学習データとして、属していない単語ペアを負の学習データとして機械学習を行う.



Step1: ドメイン毎の特徴ベクトルの決定

1. あるドメインDについて検索
2. 結果から頻出単語を抽出
   例) fruitで検索

Google	Wikipedia
fruit	fruit
tree	tree
sweet	juice
plant	sweet

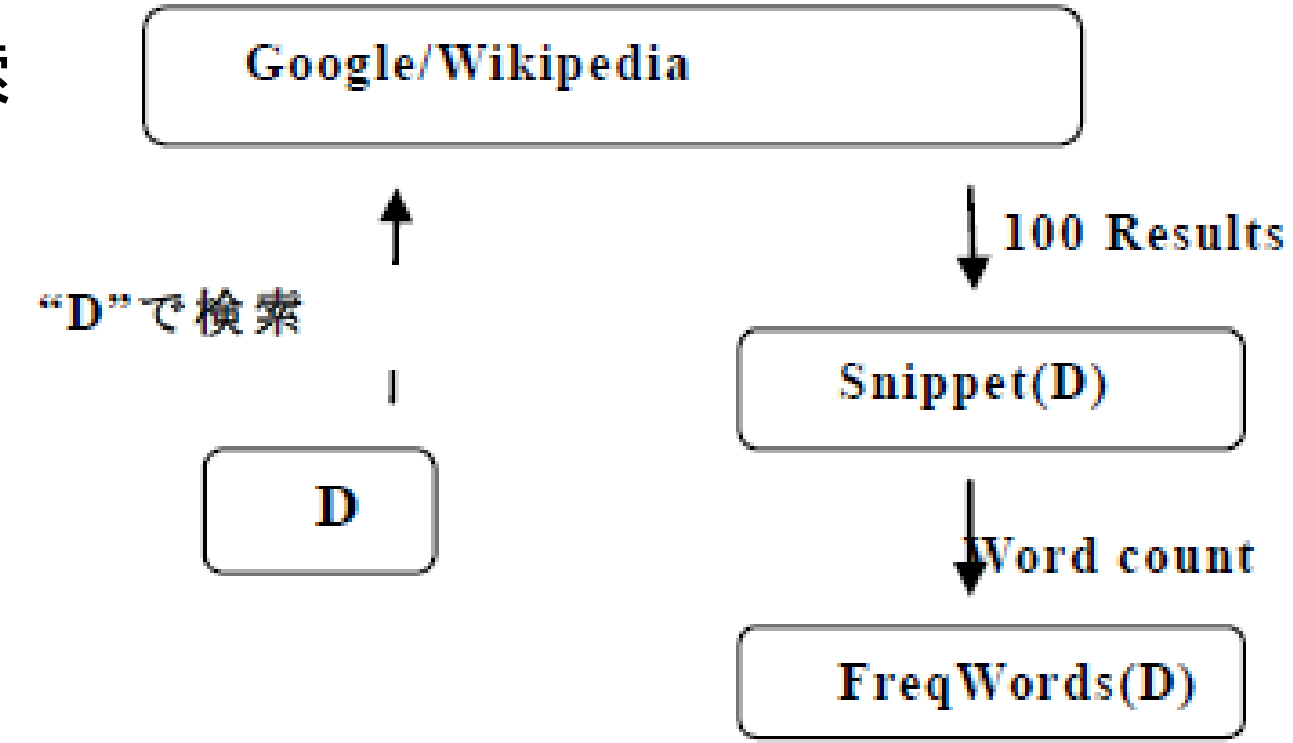


図 1 特徴ベクトルの決定

Step2: 学習データの準備

1. ドメインに属する単語ペアで検索
2. Step1で抽出された単語の頻出度を数えて学習データとする

例)apple, banana

Google		Wikipedia	
fruit	24	fruit	24
tree	4	tree	4
sweet	8	juice	14
plant	2	sweet	8

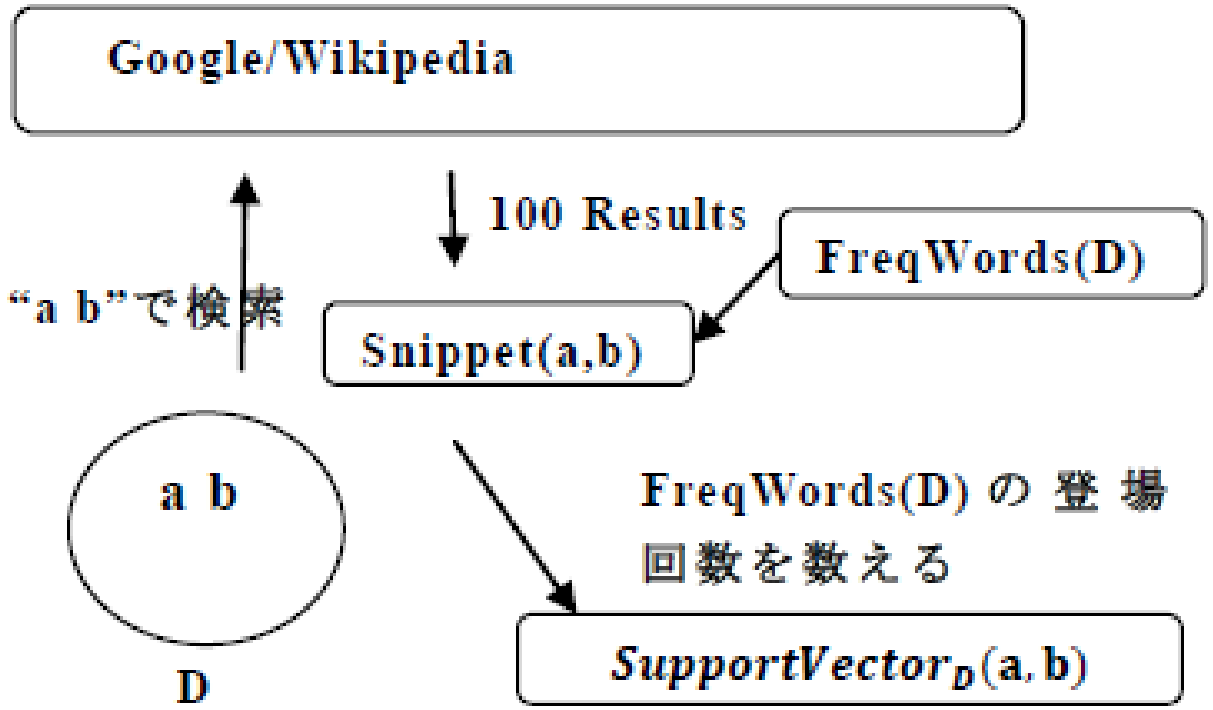


図 2 学習データの準備

Step3: SVMによる分類

1. 判別したい単語ペアも学習データと同様に特徴ベクトル化させる.
2. 学習させたSupport Vector Machineで分類させる.

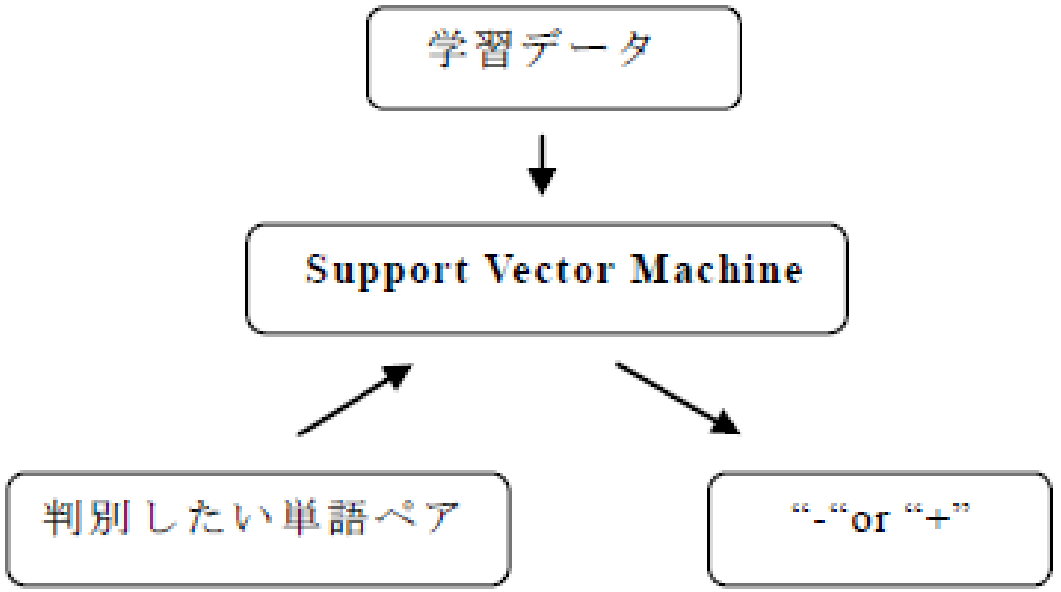


図 3 SVM による分類

Word Assignment to Domain (1)

- Economy, Geography, Medical, Communicationの4つのドメインで単語ペアを判定
- OWL-S Service Retrieval Test Collectionに収録されているこの4つのドメインに含まれているサービスネーム、インプット、　アウトプットから単語を抽出しペアとする。
- それぞれ200ペア作成しランダムに100ペア学習、100ペアをテストした.
- 例:Economyの場合Economyに関連する100ペアを正のデータとして、他の3ドメインに関連する33ペアを負のデータとする。
  (100 positive training and 99 negative training)

Word Assignment to Domain (2)

- 正しく判別できたかどうかで測定。
- 平均8割の正答率

	Test Data			
Training Data	Communication	Economy	Geography	Medical
Communication	72%	91%	99%	99%
Economy	83%	87%	85%	61%
Geography	80%	89%	77%	85%
Medical	84%	75%	70%	77%

Term Similarity and Domain (1)

- MedicalとEconomyドメインにおいて複数の単語ペアで類似度を計算した.
- サーチエンジンのヒットカウントを使ったWeb Similarity、Word Netを使ったEdgeCount、今回提案した手法の3つで類似度の計算を行う.

Customized Web PMI

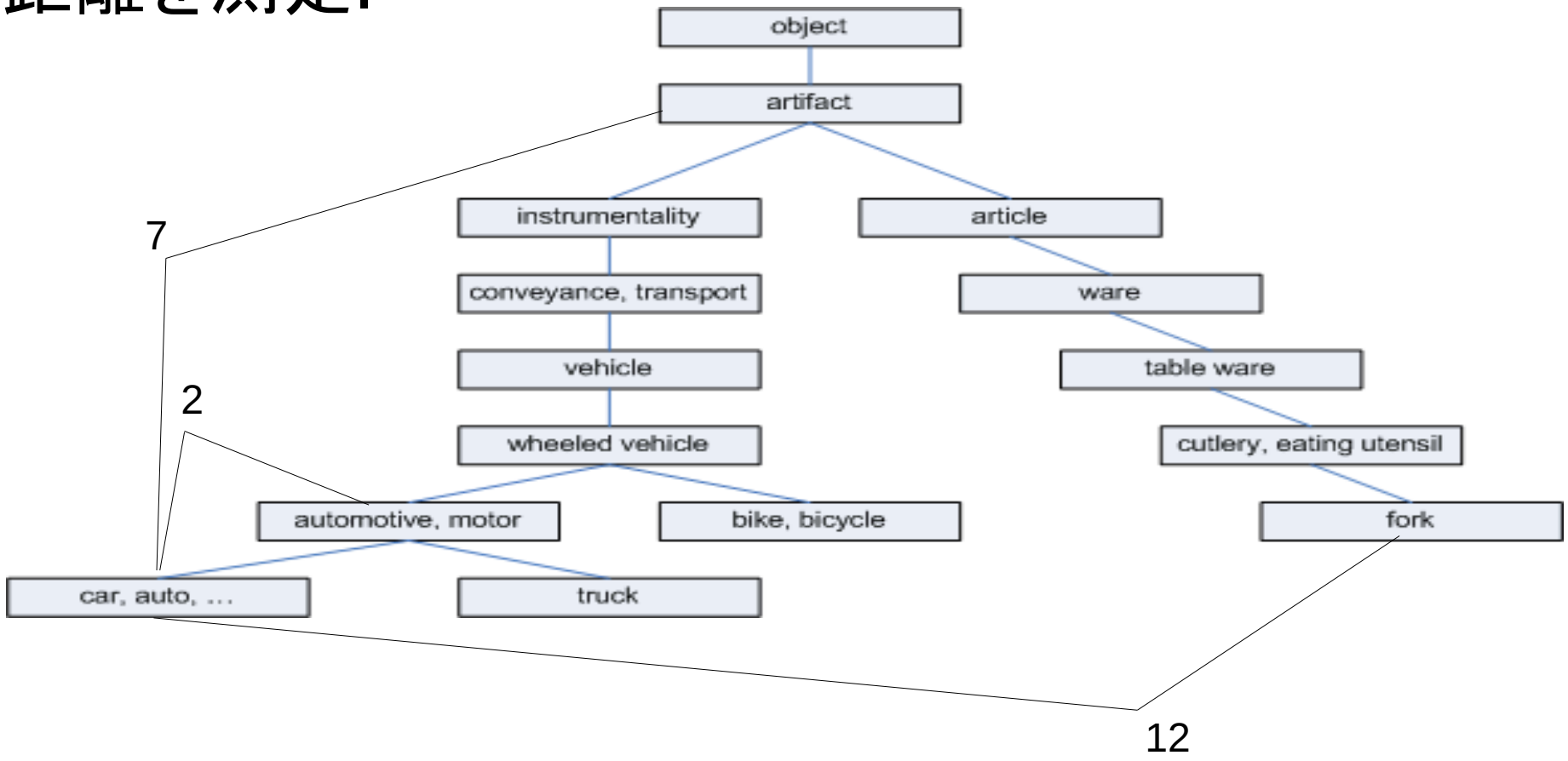
- Customized Web PMI

$$\frac{\log \frac{f(A \cap B) * N}{f(A) * f(B)}}{\log N - \log(A \cap B)}$$

- N = 30^12

Edge counting using a Word-Net

- WordNetと呼ばれる英語の意味辞書における2つの単語間の距離を測定.



Term Similarity and Domain result (1)

- WebPMIやWordNetではドメインを考慮することができないが、提案された手法で導いた値で補正することにより、求めていた結果をだすことができた。
- Economyに関係のある単語同士では1を、そうでないものは0を、Medicalでも同様になった。

Domain			WebPMI	WordNet	Proposed	Score
Medical	Disease	Clinic	0.94	0.37	1	0.77
Economy					0	0.44
Medical	Disease	Ambulance	0.84	0.48	1	0.77
Economy					0	0.44
Medical	bank	Money	1	0.85	0	0.62
Economy					1	0.95
Medical	Bank	ATM	0.95	0.62	0	0.52
Economy					1	0.86
Medical	Clinic	Money	0.83	0.29	0.04	0.39
Economy					0.16	0.43

Term Similarity and Domain result (2)

- Geographyでも同様な結果をだせた。
- MedicalとEconomyに密接に関係のある単語同士でもgeographyに関係ないものでは0になる.

Domain			WebPMI	WordNet	Proposed	Score
Geography	Location	Bank	0.86	0.73	1	0.86
	Location	ATM	0.7	0.64	1	0.78
	Location	Clinic	0.7	0.62	1	0.78
	Location	GPS	0.7	0.41	1	0.72
	Disease	Clinic	0.94	0.37	0	0.62
	Money	Bank	1.02	0.85	0	0.62

Conclusion and Future Works

- いくつかのドメインにおいて類似度を出すことができた.
- 特徴ベクトルに使われる単語の見直し、学習データの増量で精度を上げる。対応ドメインを増やす.
- 現在この手法を利用したWebサービスのクラスタリングを模索中.