3D4-5

クラス分類問題における形式概念解析を用いた近傍決定手法

Classification by Finding Neighbors in a Concept Lattice

池田 真土里　　山本 章博
Madori IKEDA　　Akihiro YAMAMOTO

京都大学 情報学研究科
Graduate School of Informatics Kyoto University

We propose a multi-class and multi-label classification method using a concept lattice. In classification, the method generates formal concepts as clusters of objects before training sets are provided, and then it selects some of the concepts for deciding a set of neighbors of an unknown object by introducing scores of the concepts. This process does not require explicit feature selection or parameters to be learned. Feature selection, which affects classification results is time-consuming in multi-class multi-label classification. We apply the method to thesaurus extension, which is a task in natural language processing and is regarded as multi-class multi-label classification. By experiments using practical thesauri and corpora, we prepare the method with the k nearest neighbor algorithm, and we show that our method works faster and generates better classification.

1. はじめに

本稿では，**形式概念**を用いて多クラス・多ラベル分類を行なう手法を提案する．対象をあらかじめクラスタリングし，未知の対象の近傍を既に生成されたクラスタから選択することでクラス分類を行なう．この手法は明示的な**特徴選択**や学習させるパラメータが不要であり，分類する対象集合に対して複数の異なる訓練集合が与えられる場合に時間の削減が期待できる．

提案手法の現実的な応用として**シソーラス拡張** [Agirre 09, Uramoto 96, Wang 06] を取り上げる．自然言語処理に欠かせない言語資源である**シソーラス**は，語の意味的な辞書であり，語には意味を表すラベルが与えられている．シソーラスの拡張とは，**未知の語**に対してシソーラスで定義されている意味ラベルを適切に与えることである．様々な種類のシソーラスが公開されており，そのほとんどにおいて膨大な意味ラベルが定義されており，各語は複数の意味ラベルが与えられている．このため，拡張するべきシソーラスは数多く存在し，シソーラス拡張は多クラス・多ラベル分類問題である．語の分類には，品詞や共起，構文構造などが語の特徴として頻繁に利用される．これらの特徴は，構文解析器などの利用により自動的に構築される**コーパス**から容易に得られる．ただし，分類に有効な語の特徴は自明でなく，コーパスは頻繁に更新される．

一般にクラス分類問題では，あらかじめ対象が持つ属性から分類に有効な**特徴の選択**を行った後に，選択された特徴を用いて未知の対象を分類する．特徴選択は分類精度に影響する重要な処理であり様々な手法が提案されている [Deng 12, Lopez 06]．特徴選択は，対象とその属性の組からなる**分類集合**と対象とそのクラスの組からなる**訓練集合**の組に対して行なわれ，多クラス・多ラベル分類問題では，適切な特徴を求めることは難しく，また時間コストが大きくなる．シソーラス拡張をクラス分類問題として解く場合，シソーラスを訓練集合，コーパスを分類集合として扱う．提案手法は，あらかじめ対象を形式概念を用いてクラスタリングした後，クラスタを選択することで対象を分類する．形式概念の生成には分類集合のみを用い，クラスタ選択は訓練集合を用いた単純な計算により行なわれる．したがって，複数の異なる訓練集合が与えられ，各訓練集合について対象集合をクラス分類する必要がある場合，時間コストの削減が期待できる．また，形式概念からなる束を漸化的に更新するアルゴリズム [Choi 06, Soldano 10, Valtchev 01] を利用することで，分類集合の更新に応じて対象のクラスタを容易に修正可能である．

本稿では，次節で形式概念と概念束の定義を示し，3 節において提案手法について述べる．4 節において，提案手法を用いてシソーラス拡張を行い，その分類精度を k 近傍法と比較する．5 節を本稿のまとめとする．

連絡先: 池田真土里，京都大学 情報学研究科，606-8501 京都市左京区吉田本町，m.ikeda@iip.ist.i.kyoto-u.ac.jp

2. 形式概念と概念束

提案手法で用いる形式概念と概念束の定義 [Ganter 99,Davey 02] を述べる．

G と M を互いに素な有限集合，$I \subseteq G \times M$ とする．G と M の要素をそれぞれ**対象**，**属性**と呼び，$(g, m) \in I$ であるとき「対象 g は属性 m を持つ」という．(G, M, I) を**コンテキスト**と呼ぶ．対象の部分集合 $A \subseteq G$ と属性の部分集合 $B \subseteq M$ に関して，

$$A^I = \{ m \in M \mid \forall g \in A, (g, m) \in I \},$$
$$B^I = \{ g \in G \mid \forall m \in B, (g, m) \in I \}$$

とする．対象集合と属性集合の組 (A, B) が $A^I = B$ かつ $A = B^I$ であるとき，(A, B) を**形式概念**と呼ぶ．また，対象 $g \in G$ について形式概念 $(\{ g \}^{II}, \{ g \}^I)$ を特に γg と記す．形式概念 $c = (A, B)$ について，A を c の**外延**と呼び $\mathrm{Ex}(c)$ と表す．任意の形式概念 c, c' について，$\mathrm{Ex}(c) \subseteq \mathrm{Ex}(c')$ であるとき $c \leq c'$ とする．コンテキスト $K = (G, M, I)$ の形式概念全体の集合に順序 $\leq$ を適用したものを**概念束** $\underline{\mathfrak{B}}(G, M, I)$ と呼ぶ．本稿では，$\underline{\mathfrak{B}}(K)$ と略記する場合がある．形式概念 $c \in \underline{\mathfrak{B}}(K)$ について，$\uparrow c = \{ c' \in \underline{\mathfrak{B}}(K) \mid c' \geq c \}$ とする．

$G_0 = \{g_1, g_2, ..., g_7\}$，$M_0 = \{m_1, m_2, ..., m_7\}$ として，コンテキストの例 $K_0 = (G_0, M_0, I_0)$ を表 1 に示す．I_0 の要素を $\times$ として記す．例えば，対象 g_4 は属性 m_2, m_4, m_6 を持つ．K_0 の概念束 $\underline{\mathfrak{B}}(K_0)$ を図 1 に示す．形式概念を円で表し，側に外延と属性集合を与えている．形式概念間の順序 $\leq$ を辺と

表 1: コンテキスト $K_0 = (G_0, M_0, I_0)$

	m_1	m_2	m_3	m_4	m_5	m_6	m_7
g_1	×	×					
g_2	×	×		×			
g_3	×	×		×			
g_4		×		×		×	
g_5		×			×	×	
g_6		×			×	×	
g_7			×		×		×

表 2: 訓練集合 $\tau_0 = (T_0, \mathcal{L}_0)$

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8
g_1	×	×						
g_2		×	×	×				
g_3				×	×	×		
g_5	×					×	×	
g_6						×	×	
g_7	×						×	×

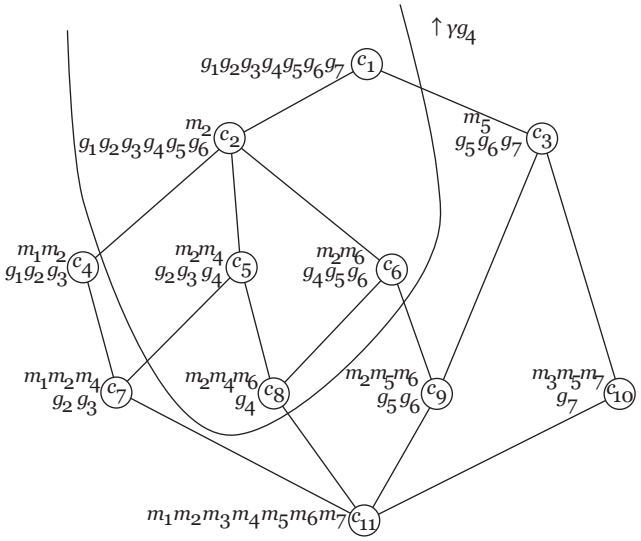

図 1: 概念束 $\underline{\mathfrak{B}}(G_0, M_0, I_0)$

して表し，上位のものを上部に記している．例えば $\gamma g_4 = c_8$，$\uparrow\gamma g_4 = \{\, c_1, c_2, c_5, c_6, c_8 \,\}$ である．

3. 概念束を用いたクラス分類手法

対象とする多クラス・多ラベル分類問題を定義した上で，提案手法について述べる．

L を G や M と素な有限集合とし，L の要素を**ラベル**と呼ぶ．各ラベル $l \in L$ は**クラス**を表す．関数 $\mathcal{L}_* : G \to 2^L$ を目標とする分類規則とし，各ラベル $l \in \mathcal{L}_*(g)$ を対象 g のラベルと呼ぶ．対象の部分集合 $T \subseteq G$ が与えられるとき，関数 $\mathcal{L} : T \to 2^L$ は $\forall g \in T.\, \mathcal{L}(g) = \mathcal{L}_*(g)$ を満たすとする．$(T, \mathcal{L})$ を**訓練集合**と呼ぶ．対象集合 G と訓練集合 $\tau = (T, \mathcal{L})$ が与えられたとき，$g \in T$ ならば対象 $g \in G$ は訓練集合 τ について**既知**，そうでなければ**未知**という．**クラス分類問題**とは，対象集合 G について訓練集合 $(T, \mathcal{L})$ を用いて関数 $\hat{\mathcal{L}} : G \to 2^L$ を決定することである．$\hat{\mathcal{L}}$ が $\forall g \in G.\, \hat{\mathcal{L}}(g) = \mathcal{L}_*(g)$ であるとき分類は**正しい**という．本稿では，$|L| \leq 2$，$|\mathcal{L}_*(g)| = 1$ とはしない多クラス・多ラベル分類問題を扱う．また，未知の対象 $u \in G \setminus T$ について $\hat{\mathcal{L}}(u)$ を推定する手法を述べる．

$T_0 = \{\, g_1, g_2, g_3, g_5, g_6, g_7 \,\}$，$\mathcal{L}_0 : T_0 \to 2^{\{\, l_1, l_2, \ldots, l_8 \,\}}$ である訓練集合の例 $\tau_0 = (T_0, \mathcal{L}_0)$ を表 2 に示す．$\hat{\mathcal{L}}_0(g_i)$ の要素を × として記す．コンテキスト $K_0 = (G_0, M_0, I_0)$ が与えられているとき，τ_0 について対象 g_4 は未知である．

提案手法は，コンテキスト形式の分類集合 (G, M, I) を用いて，与えられた訓練集合 $(T, \mathcal{L})$ について未知の対象 $u \in G \setminus T$ をクラス分類する．分類のために，コンテキストから得られる対象の特徴空間上で，既知の対象を未知の対象の**近傍**として決定し，近傍が持つラベルを未知の対象のラベルとする．具体的には，以下のステップからなる．

1. コンテキスト $K = (G, M, I)$ から概念束 $\underline{\mathfrak{B}}(K)$ を構築し，対象集合 G をクラスタリング．
2. 訓練集合 $\tau = (T, \mathcal{L})$ の未知の対象 $u \in G \setminus T$ について，近傍の**候補集合**から近傍を選択．

対象集合 G のクラスタリングは概念束 $\underline{\mathfrak{B}}(K)$ の構築により行なわれる．すなわち，概念束上の形式概念 $c \in \underline{\mathfrak{B}}(K)$ の外延 $\mathrm{Ex}(c)$ がそれぞれ対象のクラスタである．訓練集合 τ が与えられることにより決まる未知の対象 u に対して，形式概念 $c \in \uparrow\gamma u$ の外延に含まれる既知の対象 $g \in \mathrm{Ex}(c) \cap T$ を**近傍の候補**といい，$\mathrm{Ex}(c) \cap T$ を近傍の**候補集合**という．τ が与えられれば，候補集合は形式概念によって決まるため本稿ではそれらを区別しない．

近傍を決定するために次の方針を挙げる．

方針 P 未知の対象 u の近傍は互いに類似しているべきである．

方針 R 方針 P が保証される場合に限り，より多くの既知の対象を u の近傍とするべきである．

これらは，分類精度に関してそれぞれ**適合率** (precision) と**再現率** (recall) を高めることを意図している．候補集合に含まれる対象の類似性を表すために，形式概念に対して**スコア**を与える．形式概念 $c \in \underline{\mathfrak{B}}(K)$ のスコア $\sigma(c, \tau)$ を

$$\sigma(c,\tau) = \begin{cases} 0 & \text{if } |\mathrm{Ex}(c,\tau)| = 0, \\ 1 & \text{if } |\mathrm{Ex}(c,\tau)| = 1, \\ \dfrac{\sum_{i=1}^{|\mathrm{Ex}(c,\tau)|-1} \sum_{j=i+1}^{|\mathrm{Ex}(c,\tau)|} \mathrm{sim}(g_i, g_j)}{\binom{|\mathrm{Ex}(c,\tau)|}{2}} & \text{otherwise} \end{cases}$$

とする．ただし，

$$\begin{aligned} \mathrm{Ex}(c,\tau) &= \mathrm{Ex}(c) \cap T, \\ \mathrm{sim}(g_i, g_j) &= \frac{|\mathcal{L}(g_i) \cap \mathcal{L}(g_j)|}{|\mathcal{L}(g_i) \cup \mathcal{L}(g_j)|} \end{aligned}$$

とする．関数 sim は **Jaccard index** [Tan 2005] である．関数 σ は，候補集合に含まれる対象間の類似度の平均を求める．未知の対象 u について，候補集合 $c \in \uparrow\gamma u$ のうち，スコアが最大かつ外延の要素数が最大である形式概念 c の集合を $\mathrm{P}(u, \tau)$ として，近傍の集合 $\mathrm{N}(u, \tau) = \bigcup_{c \in \mathrm{P}(u,\tau)} \mathrm{Ex}(c, \tau)$ を決定する．

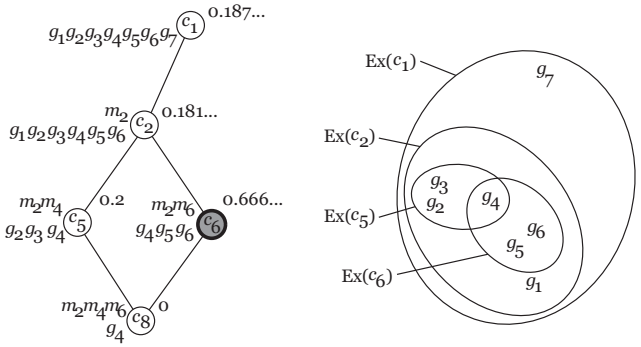

図 2: 概念束 $\underline{\mathfrak{B}}(K_0)$ 上の $\uparrow\gamma g_4$ と対象 g_4 の近傍の候補集合

$|\mathrm{P}(u,\tau)| = 1$ としないのは方針 R に基づく．最終的に，未知の対象 u に対する関数 $\hat{\mathcal{L}}$ の値を $\hat{\mathcal{L}}(u) = \bigcup_{g\in \mathrm{N}(u,\tau)} \mathcal{L}(g)$ とする．各未知の対象 u について近傍を決定するために必要な計算量は $O(|\uparrow\gamma u|M^2)$ である．M は候補集合の要素数の平均を表す．

コンテキスト K_0 と訓練集合 τ_0 の下での未知の対象 g_4 について，近傍の候補集合は図 2 右側に示すように $\mathrm{Ex}(c_1), \mathrm{Ex}(c_2), \mathrm{Ex}(c_5), \mathrm{Ex}(c_6), \mathrm{Ex}(c_8)$ である．図 2 左側に示すように，これらはそれぞれ $\uparrow \gamma g_4$ に含まれる形式概念 c_1, c_2, c_5, c_6, c_8 に対応する．各形式概念の右側にスコアを与えている．この場合，g_4 に関してスコアが最大の形式概念は c_6(図 2 内，灰色) のみであり，$\mathrm{N}(g_4, \tau_0) = \{\, g_5, g_6 \,\}$ である．結果的に，g_4 は $\hat{\mathcal{L}}_0(g_4) = \{\, l_1, l_6, l_7 \,\}$ として分類される．

4. シソーラス拡張への応用

提案手法を**シソーラス拡張**に適用する．ここでは分類する語を名詞に限定し，実験により提案手法を k **近傍法**と比較する．実験に用いたコンテキストと訓練集合について述べた後，実験結果を示す．数値は全て少数第 4 位を四捨五入している．

4.1 シソーラスとコーパス

実験では，シソーラスとして日本語 WordNet1.0 [Mok 12] と分類語意表 [NINJAL 04]，コーパスとして京都大学格フレーム 1.0 [GSK 09] と Web 日本語 N グラム第 1 版 [Kudo 04] を用いた．これら 4 つの言語資源に共通して登録されている名詞の集合を G_1 とする．日本語 WordNet と分類語意表をそれぞれ $(G_1, \mathcal{L}_{1*})$ と $(G_1, \mathcal{L}_{2*})$ とする．ただし，これらのシソーラスは意味ラベルを共有しない．表 3 にそれらの概要を示す．また，京都大学格フレームと日本語 N グラムからコンテキスト $K_1 = (G_1, M_1, I_1)$ と $K_2 = (G_1, M_2, I_2)$ をそれぞれ作成した．ただし，$I_1 \cap I_2 = \emptyset$ とする．さらに，$M_3 = M_1 \cup M_2$，$I_3 = I_1 \cup I_2$ であるコンテキスト $K_3 = (G_1, M_3, I_3)$ を用意した．表 4 に 3 つのコンテキストの概念束について示す．各未知語について近傍を求めるために必要な実際の計算量 $O(|\uparrow\gamma u|M^2)$ は，候補集合の平均数と外延の要素数の平均から概算できる．例えば k 近傍法の計算量は $O(|T|)$ であり，一般にシソーラス拡張では $|T|$ は小さくないため，本手法を用いることで時間コストを削減できることが分かる．

以下では，上記の実験に用いたデータの作成方法を述べる．$(G_1, \mathcal{L}_{1*})$ では，日本語 WordNet において「lemma」，「sense」と呼ばれる要素をそれぞれ名詞，意味ラベルとした．また $(G_1, \mathcal{L}_{2*})$ では，分類語意表で「見出し本体」,「分類番号」と呼ばれる要素をそれぞれ名詞，意味ラベルとした．

表 3: 2 種類のシソーラス

	$(G_1, \mathcal{L}_{1*})$	$(G_1, \mathcal{L}_{2*})$
名詞の数 $\lvert G_1\rvert$	7,636	7,636
意味ラベルの数	9,560	595
名詞のラベルの数の平均値	2.19	2.89

表 4: 3 種類のコーパスから得られる概念束

	$\underline{\mathfrak{B}}(K_1)$	$\underline{\mathfrak{B}}(K_2)$	$\underline{\mathfrak{B}}(K_3)$
名詞の数 $\lvert G_1\rvert$	7,636	7,636	7,636
属性の数	19,313	7,135	26,448
名詞の持つ属性数の平均値	3.85	4.70	8.55
形式概念の数	11,960	20,066	30,540
名詞に対する候補集合の平均数	2.990	14.871	18.576
形式概念の外延の要素数の平均	2.548	6.040	4.895

コンテキスト K_1 の作成に用いた京都大学格フレームは，Web 上の約 16 億文から得た**格構造**を登録したコーパスである．格構造とは，文中で述語と関係のある名詞と格助詞の組み合わせのことである．$K_1 = (G_1, M_1, I_1)$ では，M_1 の要素は格構造中の述語と格助詞の組であり，I_1 の要素がその組と名詞の関係を表す．例えば，文「犬が男に吠えている」の述語「吠える」は，格助詞「が」を伴って名詞「犬」と関係しており，同様に「に」を伴って「男」に関係している．これらの格構造から作成したコンテキストを表 5 に示す．ただし，f を格構造の頻度，n をある名詞を含む格構造全体の頻度としたとき，$0.05 \leq (f/n) \leq 0.95$ を満たす格構造のみを K_1 の作成に用いた．

日本語 N グラムは Web 上の約 200 億文から収集されたコーパスであり，日本語の文を品詞の列とみなして N グラムを定めている．このコーパスのサブセットである 4 グラムを用いてコンテキスト K_2 を作成した．$K_2 = (G_1, M_2, I_2)$ では，先頭が名詞である 4 グラムについて，名詞に続く 3 つの品詞をそれぞれその名詞の持つ属性とした．前述した例文からは「犬，が，男，に」,「男，に，吠えて，いる」という名詞を先頭に持つ 4 グラムが得られ，表 6 に示すようなコンテキストが得られる．ただし，f を 4 グラムの頻度，n をある名詞を先頭に持つ 4 グラム全体の頻度としたとき，$0.05 \leq (f/n) \leq 0.95$ を満たす 4 グラムのみを K_2 の作成に用いた．

4.2 実験結果

実験では，シソーラスの持つ名詞集合 G_1 の 10 分の 1 をランダムに選び，未知語と仮定してシソーラス拡張を行なった．コンテキストとシソーラスの各組に対して，提案手法と k 近傍法 (k-NN) を用いてシソーラス拡張を行い，適合率 (precision) と再現率 (recall) を計測した．表 7 には，シソーラス拡張をそれぞれ 10 回ずつ行なった平均値を示す．全てのシソーラス拡張において，適合率，再現率ともに提案手法の方が k 近傍法 (k-NN) より良い結果となった．

5. まとめ

本稿では，形式概念を用いて多クラス・多ラベル分類を行なう手法を提案した．分類において，あらかじめ対象のクラスタを形式概念として生成し，その後に形式概念の選択により未知の対象の近傍を決定する．この方法では，明示的な特徴選択が不要であり，学習させるパラメータなども不要である．近傍決定のために，適合度 (precision) と再現度 (recall) の向上を目

表 5: 格構造から作成したコンテキスト

	〈 吠える, が 〉	〈 吠える, に 〉
犬	×	
男		×

表 6: 4 グラムから作成したコンテキスト

	が	男	に	吠えて	いる
犬	×	×	×		
男			×	×	×

指した方針に基づいて形式概念を選択する．特に，適合度の向上を目的として，形式概念に与えたスコアを用いて近傍を選択することの有効性を示した．本手法をシソーラス拡張に適用し，k 近傍法と比べて高い精度が得られることを示した．また，その際の時間コストも現実的には削減されることを確認した．

今後は分類精度の向上が課題となる．具体的には，スコア計算方法，形式概念の選択方法，得られた近傍から未知の対象のラベルの決定方法の修正が挙げられる．表 7 より，概念束と訓練集合の組み合わせにより精度に差が見られるため，それらの組み合わせと精度の関係を解析する必要がある．

表 7: 関数 $\hat{\mathcal{L}}_1$ と $\hat{\mathcal{L}}_2$ の精度

		$(G_1, \mathcal{L}_{1*})$		$(G_1, \mathcal{L}_{2*})$	
	手法	適合率	再現率	適合率	再現率
K_1	提案手法	0.039	0.274	0.164	0.533
	1-NN	0.026	0.024	0.103	0.103
	5-NN	0.007	0.036	0.031	0.150
	10-NN	0.004	0.038	0.016	0.169
K_2	提案手法	0.007	0.079	0.028	0.248
	1-NN	0.007	0.007	0.027	0.027
	5-NN	0.002	0.013	0.014	0.070
	10-NN	0.002	0.018	0.010	0.100
K_3	提案手法	0.030	0.072	0.132	0.250
	1-NN	0.009	0.009	0.039	0.039
	5-NN	0.004	0.018	0.017	0.085
	10-NN	0.002	0.024	0.011	0.116

参考文献

[Agirre 09] Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching Very Large Ontologies Using the WWW. Proc. of ECAI'00 Workshop “Ontology Learning” (2000)

[Choi 06] Choi, V., Huang, Y.: Faster Algorithms for Constructing a Galois Lattice, Enumerating All Maximal Bipartite Cliques and Closed Frequent Sets. SIAM Conference on Discrete Mathematics (2006)

[Davey 02] Davey, B., A., Priestly, H., A.: Introduction to Lattice and Order. Cambridge University Press (2002)

[Deng 12] Deng, H., Runger, G.: Feature Selection via Regularized Trees. Proc. of IJCNN'12, pp. 1–8. (2012)

[Ganter 99] Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag (1999)

[GSK 09] 言語資源協会, `http://www.gsk.or.jp` (2009)

[Kudo 04] Kudo, T., Kazawa, H.: Web Japanese N-gram Version 1, Gengo Shigen Kyokai (2004)

[NINJAL 04] 国立国語研究所, `http://www.ninjal.ac.jp/archives/goihyo` (2004)

[Lopez 06] Lopez, F., G., Torres, M., G., Melian, B., Perez, J., A., M., Moreno-Vega, J., M.: Solving Feature Subset Selection Problem by a Parallel Scatter Search. European Journal of Operational Research, vol. 169, no. 2, pp. 477–489 (2006)

[Mok 12] Mok, S., W., H., Gao, H., E., Bond, F.: Using Wordnet to Predict Numeral Classifiers in Chinese and Japanese. Proc. of GWC'12 (2012)

[Soldano 10] Soldano, H., Ventos, V., Champesme, M., Forge, D.: Incremental Construction of Alpha Lattices and Association Rules. Proc. of KES'10, pp. 351–360. Springer (2010)

[Tan 2005] Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley (2005)

[Uramoto 96] Uramoto, N.: Positioning Unknown Words in a Thesaurus by Using Information Extracted from a Corpus. Proc. of COLING'96, vol. 2, pp. 956–961. Association for Computational Linguistics (1996)

[Valtchev 01] Valtchev, P., Missaoui, R.: Building Concept (Galois) Lattices from Parts: Generalizing the Incremental Methods. Proc. of ICCS'01, pp. 290–303. Springer (2001)

[Wang 06] Wang, J., Ge, N.: Automatic Feature Thesaurus Enrichment: Extracting Generic Terms From Digital Gazetteer. Proc. of JCDL'06, pp. 326–333. IEEE (2006)