computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

MODEL OF SEMANTIC CONTEXT OF LEXEMES IN THE TEXT MINING ALGORITHMS

Bohdan M. Pavlyshenko

Ivan Franko Lviv National University,
Drahomanov Str. 50, Lviv, 79005 Ukraine, e-mail:pavlsh@yahoo.com

Abstract: *The model of semantic context of lexemes which represent the structure semantic configuration of lexems corpus of text arrays has been proposed. It is shown that partially ordered set of semantic concepts are formed in the lexem semantic context. Concepts' intents are defined by semantic fields, concepts extents – by lexems.*

Key words: *Formal Concepts Analysis, Semantic Context, Semantic Fields, Text Mining.*

Formal concept analysis is one of components of modern data mining methods [1, 2, 3, 4]. In this theory the formal contexts are analysed using the algebraic lattices. The use of model of lexeme semantic context can be effective in the text mining algorithims. Let consider the theoretical model of text documents, dictionary of lexemes and semantic fields. Let some lexeme dictionary exist

$$W = \{ w_i \mid i = 1,2..., N_w \} \quad (1)$$

Texts documents can be described as

$$D = \{ d_j \mid j = 1,2..., N_d \} \quad (2)$$

Introduce semantic fields set

$$S = \{ s_k \mid k = 1,2..., N_s \} \quad (3)$$

Mapping of lexemes of text dictionary to semantic fields set can be represented as

$$U_{ws} : w_i \to s_k, \ i = 1,2..., N_w; k = 1,2..., N_s \quad (4)$$

The set of lexems of semantic field s_k can be written as

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\to} s_k, i = 1,2..., N_w \right\} \quad (5)$$

The matrix of semantic attributes can be written as

$$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1, j=1}^{N_s, N_d} \quad (6)$$

where p_{kj}^{sd} – the frequency of semantic field s_k in the lexems set of document

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t}. \quad (7)$$

The vector

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right) \quad (8)$$

represents the document d_j in N_s-dimensional space. Let define semantic context of lexems as triple

$$\mathrm{K}_s = (W, S, I), \quad (9)$$

where W – the set of lexems of texts dictionary, S – the set of semantic fields, I – relation:

$$I \subseteq W \times S, \ \ I = \left\{ \langle w_i, s_k \rangle \right\} \quad (10)$$

Pair $\langle w_i, s_k \rangle$ means that lexeme w_i belongs to the semantic field s_k. Let define the lattice of semantic concepts of lexems. For $Extent \subseteq W$, $Intent \subseteq S$ let define the following mappings

$$\begin{aligned} Extent' &= \{ s \in S \mid w \in Extent : wIs \} \\ Intent' &= \{ w \in W \mid s \in Intent : wIs \} \end{aligned} \quad (11)$$

Let consider the semantic concept as a pair

$$Concept = (Extent, Intent) \quad (12)$$

where $Extent \subseteq W$, $Intent \subseteq S$ are with the following condictions

$$\begin{aligned} Extent' &= Intent \\ Intent' &= Extent \end{aligned} \quad (13)$$

The set $Extent$ is called extent, $Intent$ – intent of semantic concept $Concept$. Partially ordered set appears in the semantic context of lexemes

$$\begin{aligned} \Psi(W, S, I) &= \{ Concept_m \,|\, m = 1,2,...N_{ct} \}, \\ Concept_m &= (Extent_m,\ Intent_m), \end{aligned} \quad (14)$$

where N_{ct} – amount of concepts. The concepts lattice can be presented by Hasse diagram. Let consider an example when lexemes $[L1, L2, L3, L4]$ belong to semantic fields $[S1, S2, S3, S4]$.

The example of Hasse diagram of context in a case of univalued correspondence between lexemes and semantic fields, when different lexemes belong to different semantic fields, is showed in Fig. 1. The example of Hasse diagram of context in the case of multivalued correspondence between lexemes and semantic fields, when some lexems can belong to different fields at the same time, is showed in Fig. 2. Frequency characteristics of extents of received semantic concepts give new components for vector (8) and generate additional subspace for quantitative analysis.

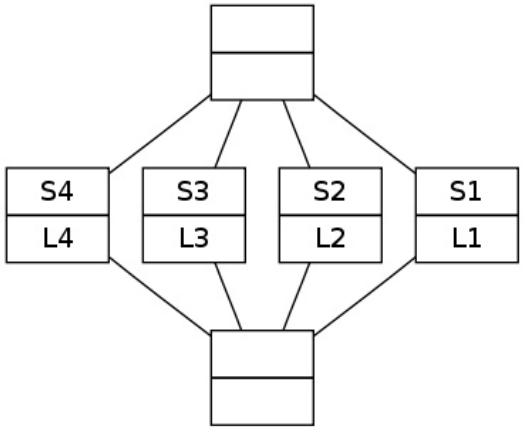

Fig. 1 – Example of Hasse diagram in case of univalued correspondence between lexemes and semantic fields

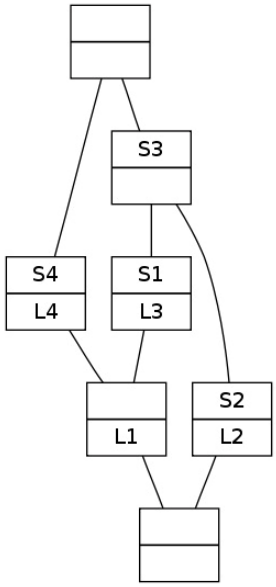

Fig. 2 – Example of Hasse diagram in case of multivalued correspondence between lexemes and semantic fields

CONCLUSION

The considered model of semantic context of lexemes represents the structural semantic configuration of lexems corpus of text arrays. It is shown that partially ordered set of semantic concepts are formed in the lexeme semantic context. Concepts' intents are defined by semantic fields, concepts extents – by lexemes. Using the model suggested in this work gives the ability to detect new subsets of semantic groups of lexemes for forming the vector space for texts' arrays analysis.

REFERENCES

[1] B. Ganter, R. Will, *Formal Concept Analysis: Mathematical Foundations*, Springer, 1999.

[2] S.O. Kuznetsov, S.A. Obiedkov, Comparing performance of algorithms for generating concept lattices, *Journal of Experimental and Theoretical Artificial Intelligence*, 14 (2002). – pp. 189-216.

[3] P. Cimiano, A. Hotho, S. Staab, Learning concept hierarchies from text corpora, using formal concept analysis, *Journal of Artificial Intelligence Research*, 24 (2005). – pp. 305-339.

[4] Abderrahim El Qadi, Driss Aboutajedine, Yassine Ennouary, Formal concept analysis for information retrieval, *International Journal of Computer Science and Information Security (IJCSIS)*, 7 (2) (2010).