arXiv:1904.08010v1 [cs.IR] 16 Apr 2019

How to define *co-occurrence* in different domains of study?

Mathieu Roche [1,2]

[1] Cirad, TETIS, F-34398 Montpellier, France
[2] TETIS, Univ. Montpellier, AgroParisTech, Cirad, CNRS, Irstea, Montpellier, France
mathieu.roche@cirad.fr
http://textmining.biz/Staff/Roche

Abstract. This position paper presents a comparative study of *co-occurrences*. Some similarities and differences in the definition exist depending on the research domain (e.g. linguistics, NLP, computer science). This paper discusses these points, and deals with the methodological aspects in order to identify *co-occurrences* in a multidisciplinary paradigm.

Keywords: co-occurrence, collocation, phrase, n-gram, skyp-n-gram, association rule, sequential pattern

1 Introduction

Determining *co-occurrences* in corpora is challenging for different applications such as classification, translation, terminology building, etc. More generally, *co-occurrences* can be identified with all types of data, e.g. databases [1], texts [2], images [3], music [4], video [5], etc.

The *co-occurrence* concept has different definitions depending on the research domain (i.e. linguistics, NLP, computer science, biology, etc.). This position paper reviews the main definitions in the literature and discusses similarities and differences according to the domains. This type of study can be crucial in the context of data science, which is geared towards developing a multidisciplinary paradigm for data processing and analysis, especially textual data.

Here the *co-occurrence* concept related to textual data is discussed. Note that before their validation by an expert, co-occurrences of words are often considered as *candidate terms*.

First, Section 2 of this paper details the different definitions of *co-occurrence* according to the studied domains. Section 3 discusses and compares these different aspects based on their intrinsic definition but also on the associated methodologies in order to identify them. Finally, Section 4 lists some perspectives.

2 *Co-occurrence* in a multidisciplinary context

2.1 Linguistic viewpoint

In linguistics, one notion that is widely used to define the term is called *lexical unit* [6] and *polylexical expression* [7]. The latter represents a set of words having an autonomous existence, which is also called *multi-word expression* [8].

In addition, several linguistics studies use the *collocation* notion. [9] gives two properties defining a collocation. First, collocation is defined as a group of words having an overall meaning that is deducible from the units (words). For example, *climate change* is considered as a collocation because the overall meaning of this group of words can be deduced from both words *climate* and *change*. On the other hand, the expression *to rain cats and dogs* is not a collocation because its meaning cannot be deduced from each of the words; this is called a *fixed expression* or an *idiom*.

A second property is added by [9] to define a collocation. The meaning of the words that make up the collocation must be limited. For example, *buy a dog* is not a collocation because the meaning of *buy* is not limited.

2.2 NLP viewpoint

In the natural language processing (NLP) domain, the *co-occurrence* notion refers to the general phenomenon where words are present together in the same context. More precisely, several principles are used that take contextual criteria into account.

First, the terms or phrases [10,11] can respect syntactic patterns (e.g. adjective noun, noun noun, noun preposition noun, etc.). Some examples of extracted phrases (i.e. *syntactic co-occurrences*) are given in Table 1.

In addition, methods without linguistic filtering are also conventionally used in the NLP domain by extracting n-grams of words (i.e. *lexical co-occurrences*) [12,13]. n-grams are contiguous sequences of n words extracted from a given sequence of text (e.g. the bi-grams[3] $x\ y$ and $y\ z$ are associated with the text $x\ y\ z$). n-grams that allow gaps are called skip-n-grams (e.g. the skip-bi-grams $x\ y$, $x\ z$, $y\ z$ are related to the text $x\ y\ z$). Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships [14]. Some examples of n-grams and skip-n-grams are given in Table 1.

After summarizing the term notion in the NLP domain, the following section discusses these aspects in the computer science context, particularly in data

[3] n-grams with $n = 2$.

mining. Note that the NLP domain may be considered as being located at the linguistics and computer science interface.

Sentence (input) *With climate change the water cycle is expected to undergo significant change.*	
Candidates (output)	
Phrases (noun noun, adjective-noun)	*climate change* *water cycle, significant change*
bi-grams of words	*With climate, climate change, change the, the water,* *water cycle, cycle is, is expected, expected to,* *to undergo, undergo significant, significant change*
2-skip-bi-grams	*With climate, With change, With the,* *climate change, climate the, climate water,* *change the, change water, change cycle,* *the water, the cycle, the is,* *water cycle, water is, water expected,* *cycle is, cycle expected, cycle to,* *is expected, is to, is undergo,* *expected to, expected undergo, expected significant,* *to undergo, to significant, to change,* *undergo significant, undergo change,* *significant change*

Table 1. Examples of candidates extracted with different NLP techniques.

2.3 Computer science viewpoint

In the data mining domain, co-occurring items are called *association rules* [15,16] and they could be candidates for construction or enrichment of terminologies [17].

In the data mining context, the list of items corresponds to the set of available articles. With textual data, items may represent the words present in sentences, paragraphs, or documents [18,19]. A transaction is a set of items. A set of transactions is a learning set used to determine association rules.

Some extensions of association rules are called *sequential patterns*. They take into account a certain order of extracted elements [20,21] with an enriched representation related to textual data as follows:

– *objects* represent texts or pieces of texts,
– *items* are the words of a text,
– *itemsets* represent sets of words present together within a sentence, paragraph or document,
– *dates* highlight the order of sentences within a text.

There are several algorithms for discovering association rules and sequential patterns. One of the most popular is Apriori, which is used to extract frequent itemsets from large databases. The Apriori algorithm [15] finds frequent itemsets where k-itemsets are used to generate $k+1$-itemsets.

Association rules and sequential patterns of words are often used in text mining for different applications, e.g. terminology enrichment [17], association of concept instances [22,19], classification [20,21], etc.

3 Discussion: comparative study of definitions and approaches

This section proposes a comparison of : (i) *co-occurrence* definitions (see Section 3.1), (ii) automatic methods in order to identify them (see Section 3.2). This section highlights some similarities and differences between domains.

3.1 *Co-occurrence* extraction

The general definition of *co-occurrence* is finally close to *association rules* in data mining domain. Note that the integration of windows[4] in the association rule or sequential pattern extraction process enables us to have similarity with skip-n-gram extraction.

The integration of syntactic criteria makes it possible to extract more relevant candidate terms (see Table 1). Such information is typically taken into account in NLP to extract terms from general or specialized domains [23,24,25,26].

Table 1 highlights relevant terms extracted using linguistic patterns (e.g. *climate change, water cycle, significant change*). The use of linguistic patterns tends to improve precision values. Generally other methods such as skip-bi-grams return lower precision, i.e. many extracted candidates are irrelevant (e.g. *climate the*). But this kind of method enables extraction of some relevant terms not found with linguistic patterns (e.g. *cycle expected*); then the recall can be improved.

Table 2 presents research domains related to different types of candidates, i.e. collocations, polylexical expressions, phrases, n-grams, association rules, sequential patterns.

[4] Association Rule with Time-Windows (ARTW) [16].

Table 3 summarizes the main criteria described in the literature. Note that the extraction is more flexible and automatic when there are fewer criteria. In this table, two types of information are associated with the different criteria. The first one (marked with ✓) designates the characteristics given by the *co-occurrence* definitions. The second type of information (marked with ★) represents characteristics that are implemented in many extensions of the state-of-the-art.

Definitions	Domains
Collocations	L
Polylexical expressions	L + NLP
Phrases	NLP
n-grams	NLP + CS
Association rules	CS
Sequential patterns	CS

Table 2. Summary of the main domains associated with expressions (L: linguistics, NLP: natural language processing, CS: computer science).

	Ordered sequences	Sequences with gaps	Morpho-syntactic information	Semantic information
Collocations	✓		✓	★
Polylexical expressions	✓		✓	
Phrases	✓		✓	
n-grams	✓	★		
Association rules		✓		
Sequential patterns	✓	✓		

Table 3. Summary of the main criteria associated with *co-occurrence* identification. ✓ represents the respect of the criterion by definition. ★ is present when extensions are currently used in the state-of-the-art.

Table 3 shows that the semantic criterion is seldom associated with *co-occurrence* definitions. This criterion is however taken into account in linguistics. For example, semantic aspects are taken into account in several studies [27,28,29]. In this context [29] introduced *lexical functions* rely on semantic criteria to define the relationships between collocation units. For instance, a given relation can be expressed in various ways between the arguments and their values, like *Centr (the center, culmination of)* that returns different meanings[5]:

– *Centr(crisis) = the peak*

[5] http://people.brandeis.edu/∼smalamud/ling130/lex_functions.pdf

– *Centr(desert) = the heart*
– *Centr(forest) = the thick*
– *Centr(glory) = summit*
– *Centr(life) = prime*

In the data mining domain, semantic information is used in two main directions. The first one involves filtering the results if they respect certain semantic information (e.g. phrases or patterns where a word is an instance of a semantic resource). Other methods involve semantic resources in the knowledge discovery process, i.e. the extraction is driven by semantic information [22].

In recent studies in the NLP domain, the semantic aspects are based on word embedding, which provides a dense representation of words and their relative meanings [30,31].

Finally, note that several types of *co-occurrence* are often used in different domains. For example, polylexical expressions are commonly used in NLP and also in linguistics. In addition, n-grams is currently used in NLP and computer science domains. For example, n-grams of words are often used to build terminologies (NLP domain) but also as features for machine learning algorithms (computer science domain) [13].

Table 4 summarizes the main types of criteria (i.e. statistic, morpho-syntactic, and semantic) used for extracting *co-occurrences* according to the research domains considered in this paper.

	Statistic information	Morpho-syntactic information	Semantic information
Linguistics		✓	★
NLP	✓	✓	★
Data mining	✓	★	★

Table 4. Summary of the main criteria associated with research domains. ✓ represents the respect of the criterion for extracting *co-occurrences* from textual data. ★ is present when extensions are currently used in the state-of-the-art.

After presenting the characteristics associated with the *co-occurrence* notion in a multidisciplinary context, the following section compares the methodological viewpoints to identify these elements according to the domains.

3.2 Ranking of *co-occurrences*

Co-occurrence identification by automatic systems is generally based on the use of quality measures and/or algorithms. This section provides two illustrative ex-

amples that show similarities between approaches according the domains.

Mutual Information and *Lift measure*

First the use of specific statistical measures from different domains is highlighted. This paragraph focuses on the study of Mutual Information (MI). This measure is often used in the NLP domain to measure the association between words [32]. MI (see formula (1)) compares the probability of observing x and y together (joint probability) with the probability of observing x and y independently (chance) [32].

$$I(x) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

In general, word probabilities $P(x)$ and $P(y)$ correspond to the number of observations of x and y in a corpus, normalized by the size of the corpus. Some extensions of MI are also proposed. The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval) described in [33] queries the Web via the AltaVista search engine to determine appropriate synonyms for a given query. For a given word, denoted x, PMI-IR chooses a synonym among a given list. These selected terms, denoted y_i, $i \in [1, n]$, correspond to TOEFL questions. The aim is to compute the y_i synonym that gives the best score. To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present. Turney's formula is given below (2): It is one of the basic measures used in [33]. It is inspired from MI described in [32]. With this formula (2), the proportion of documents containing both x and y_i (within a 10 word window) is calculated, and compared with the number of documents containing the word y_i. The higher this proportion, the more x and y_i are seen as synonyms.

$$score(y_i) = \frac{nb(x\ NEAR\ y_i)}{nb(y_i)} \tag{2}$$

- $nb(x)$ computes the number of documents containing the word x (i.e. nb corresponds to number of webpages returned by search engines),
- $NEAR$ (used in the 'advanced research' field of AltaVista) is an operator that identifies if two words are present in a 10 word wide window.

This kind of web mining approach is also used in many NLP applications, e.g. (i) computing the relationship between *host* and *clinical sign* for an epidemiology surveillance system [34], (ii) computing the dependency of words of acronym definitions for word-sense disambiguation tasks [35].

The probabilities are generally symmetric (i.e. $P(x, y) = P(y, x)$), while the original MI measure is also symmetric. But the association ratio applied in the NLP domain is not symmetric, i.e. the occurrence number of pairs of words "x y" and "y x" generally differ. Moreover the meaning and relevance of phrases should differ according to the word order in a text, e.g. *first lady* and *lady first*.

Finally, MI is very close to the *lift* measure [36,37,38] in data mining. This measure identifies relevant association rules (see formula (3)). The lift measure evaluates the relevance of co-occurrences only (not implication) and how x and y are independent [38].

$$lift(x \rightarrow y) = \frac{conf(x \rightarrow y)}{sup(y)} \tag{3}$$

This measure is based on both *confidence* and *support* criteria, which in turn are based on association rule ($x \rightarrow y$) identification. Support is an indication of how frequently the itemset appears in the dataset. Confidence is a standard measure that estimates the probability of observing y given x (see formula 4).

$$conf(x \rightarrow y) = \frac{sup(x \cup y)}{sup(x)} \tag{4}$$

Note that other quality measures of the data mining domain, such as *Least contradiction* or *Conviction* [39], could be tailored to deal with textual data.

C-value and *closed itemset*

Another example is the methodological similarities associated with different approaches. For example, the C-value approach [40] used in the NLP domain [24,23] favors terms that do not appear to a significant extent in longer terms. For example, in a specialized corpus related to ophthalmology, [40] show that a more general term such as *soft contact* is irrelevant, whereas a longer and therefore more specific term such as *soft contact lens* is relevant. This kind of measure is particularly relevant in the biology domain [24,23].

In addition, in the computer science domain (i.e. data mining), the notion of *closed itemset* is finally very close to the C-value approach. In this context, a frequent itemset is considered as closed if none of its supersets[6] has the same support (i.e. frequency).

This section and both illustrative examples confirm the importance of having a real multidisciplinary viewpoint on the methodological aspects, in order to build scientific bridges and thus contribute to the development of the emerging data science domain.

[6] A superset is defined with respect to another itemset, for example {M1, M2, M3} is a superset of {M1, M2}. B is superset of A if card(A) $<$ card(B) and A $\subset$ B.

4 Conclusion and Future Work

This position paper proposes a discussion on similarities as well as differences in the definition of *co-occurrence* according to research domains (i.e. linguistics, NLP, computer science). The aim of this position paper is to show the bridges that exist between different domains.

In addition, this paper highlights some similarities in the methodologies used in order to identify *co-occurrences* in different domains. We could extend the discussion to other domains. For example, methodological transfers are currently applied between bioinformatics and NLP. For example, the use of edition measures (e.g. Levenshtein distance) for sequence alignment tasks (bioinformatics) *v.s.* string comparison (NLP).

Acknowledgments

This work is funded by the SONGES project (Occitanie and FEDER) – Heterogeneous Data Science (`http://textmining.biz/Projects/Songes`).

References

1. Cao, H., Hripcsak, G., Markatou, M.: A statistical methodology for analyzing co-occurrence data from a large sample. Journal of Biomedical Informatics **40**(3) (2007) 343 – 352
2. Roche, M., Azé, J., Matte-Tailliez, O., Kodratoff, Y.: Mining texts by association rules discovery in a technical corpus. In: Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004. (2004) 89–98
3. Verma, M., Raman, B., Murala, S.: Local extrema co-occurrence pattern for color and texture image retrieval. Neurocomput. **165**(C) (October 2015) 255–269
4. Ghosal, A., Chakraborty, R., Dhara, B.C., Saha, S.K. In: Song Classification: Classical and Non-classical Discrimination Using MFCC Co-occurrence Based Features. Springer Berlin Heidelberg, Berlin, Heidelberg (2011) 179–185
5. Jeon, H.H., Basso, A., Driessen, P.F. In: Camera Motion Detection in Video Sequences Using Motion Cooccurrences. Springer Berlin Heidelberg, Berlin, Heidelberg (2005) 524–534
6. Lederer, C.: La notion d'unité lexicale et l'enseignement du lexique. The French Review **43**(1) (1969) 96–98
7. Gross, G.: Les expressions figées en français. Ophrys (1996)
8. Sag, I.A., Baldwin, T., Bond, F., Copestake, A.A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '02, London, UK, UK, Springer-Verlag (2002) 1–15

9. Clas, A.: Collocations et langues de spécialité. Meta **39**(4) (1994) 576–580
10. Bourigault, D.: Surface grammatical analysis for the extraction of terminological noun phrases. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 3. COLING '92, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 977–981
11. Daille, B., Gaussier, E., Langé, J.M.: Towards automatic extraction of monolingual and bilingual terminology. In: Proceedings of the 15th Conference on Computational Linguistics - Volume 1. COLING '94, Stroudsburg, PA, USA, Association for Computational Linguistics (1994) 515–521
12. Massung, S., Zhai, C.: Non-native text analysis: A survey. Natural Language Engineering **22**(2) (2016) 163–186
13. Tungthamthiti, P., Shirai, K., Mohd, M. In: Recognition of sarcasm in tweets based on concept level sentiment analysis and supervised learning approaches. Faculty of Pharmaceutical Sciences, Chulalongkorn University (2014) 404–413
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13, USA, Curran Associates Inc. (2013) 3111–3119
15. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1994) 487–499
16. Yin, Y., Kaku, I., Tang, J., Zhu, J. In: Association Rules Mining in Inventory Database. Springer London, London (2011) 9–23
17. Di-Jorio, L., Bringay, S., Fiot, C., Laurent, A., Teisseire, M.: Sequential patterns for maintaining ontologies over time. In: On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II. (2008) 1385–1403
18. Amir, A., Aumann, Y., Feldman, R., Fresko, M.: Maximal association rules: A tool for mining associations in text. Journal of Intelligent Information Systems **25**(3) (Nov 2005) 333–345
19. Rabatel, J., Lin, Y., Pitarch, Y., Saneifar, H., Serp, C., Roche, M., Laurent, A.: Visualisation des motifs séquentiels extraits à partir d'un corpus en ancien français. In: Extraction et gestion des connaissances (EGC'2008), Actes des 8èmes journées Extraction et Gestion des Connaissances, Sophia-Antipolis, France, 29 janvier au 1er février 2008, 2 Volumes. (2008) 237–238
20. Jaillet, S., Laurent, A., Teisseire, M.: Sequential patterns for text categorization. Intell. Data Anal. **10**(3) (May 2006) 199–214
21. Serp, C., Laurent, A., Roche, M., Teisseire, M.: La quête du graal et la réalité numérique. Corpus **7** (2008)
22. Berrahou, S.L., Buche, P., Dibie, J., Roche, M.: Xart: Discovery of correlated arguments of n-ary relations in text. Expert Systems with Applications **73**(Supplement C) (2017) 115 – 124
23. Jiang, M., Denny, J.C., Tang, B., Cao, H., Xu, H.: Extracting semantic lexicons from discharge summaries using machine learning and the c-value method. In: AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, Illinois, USA, November 3-7, 2012. (2012)
24. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: Biomedical term extraction: Overview and a new methodology. Information Retrieval Journal **19**(1-2) (April 2016) 59–99

25. Nenadić, G., Spasić, I., Ananiadou, S.: Terminology-driven mining of biomedical literature. In: Proceedings of the 2003 ACM Symposium on Applied Computing. SAC '03, New York, NY, USA, ACM (2003) 83–87
26. Roche, M., Teisseire, M., Shrivastava, G.: Valorcarn-TETIS: Terms extracted with Biotex [dataset]. CIRAD Dataverse (2017)
27. Heid, U.: Towards a corpus-based dictionary of german noun-verb collocations. In: Proceedings of the Euralex International Congress. (1998) 301–312
28. Laurens, M.: La description des collocations et leur traitement dans les dictionnaires. Romaneske **4** (1999) 44–51
29. Mel'čuk, I.A., Arbatchewsky-Jumarie, N., Elnitsky, L., Lessard, A.: Dictionnaire explicatif et combinatoire du francais contemporain. Presses de l'Université de Montréal, Montréal, Canada (1984, 1988, 1992, 1999) Volume 1, 2, 3, 4.
30. Ganguly, D., Roy, D., Mitra, M., Jones, G.J.: Word embedding based generalized language model for information retrieval. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15, New York, NY, USA, ACM (2015) 795–798
31. Zamani, H., Croft, W.B.: Relevance-based word embedding. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '17, New York, NY, USA, ACM (2017) 505–514
32. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1) (March 1990) 22–29
33. Turney, P.D.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of the 12th European Conference on Machine Learning. EMCL '01, London, UK, UK, Springer-Verlag (2001) 491–502
34. Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B.: Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. International Journal of Agricultural and Environmental Information Systems **7**(3) (2016) 1–20
35. Roche, M., Prince, V.: A web-mining approach to disambiguate biomedical acronym expansions. Informatica (Slovenia) **34**(2) (2010) 243–253
36. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: Generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. SIGMOD '97, New York, NY, USA, ACM (1997) 265–276
37. Ventura, S., Luna, J.M. In: Quality Measures in Pattern Mining. Springer International Publishing, Cham (2016) 27–44
38. Azevedo, P.J., Jorge, A.M.: Comparing rule measures for predictive association rules. In: Proceedings of the 18th European Conference on Machine Learning. ECML '07, Berlin, Heidelberg, Springer-Verlag (2007) 510–517
39. Lallich, S., Teytaud, O., Prudhomme, E. In: Association Rule Interestingness: Measure and Statistical Validation. Springer Berlin Heidelberg, Berlin, Heidelberg (2007) 251–275
40. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries **3**(2) (Aug 2000) 115–130