Quotations, Relevance and Time Depth: Medieval Arabic Literature in Grids and Networks

Petr Zemánek
Institute of Comparative Linguistics
Charles University, Prague
Czech Republic
petr.zemanek@ff.cuni.cz

Jiří Milička
Institute of Comparative Linguistics
Charles University, Prague
Czech Republic
jiri@milicka.cz

Abstract

This contribution deals with the use of quotations (repeated n-grams) in the works of medieval Arabic literature. The analysis is based on a 420 millions of words historical corpus of Arabic. Based on repeated quotations from work to work, a network is constructed and used for interpretation of various aspects of Arabic literature. Two short case studies are presented, concentrating on the centrality and relevance of individual works, and the analysis of a time depth and resulting impact of a given work in various periods.

1 Quotations and Their Definition

The relevance of individual works in a given literature and the time depth of such relevance are of interest for many reasons. There are many methods that can reveal such relevance.

The current contribution is based on quotation extraction. Quotations, both covert and overt, both from written and oral sources, belong to constitutive features of medieval Arabic literature.

There are genres which heavily depend on establishing credible links among sources, especially the oral ones, where a trusty chain of tradents is crucial for the claims that such chains accompany. Other links may point to the importance of a given work (or its part) and may uncover previously unseen relations within a given literature or a given genre/register, or reveal connections among genres/registers within a given literature. As such, the results are interesting in a wide research range, from linguists or literature theorists to authors interested in the interactions of various subsets of a given literature.

The research on quotations, their extraction and detection is rich in the NLP, but the algorihms used are based mainly on the quotation-marker recognition, e.g. Pareti et al. (2013), Pouliquen et al. (2007) and Fernandes et al. (2011), or on the metadata procesing (e.g. Shi et al. 2010), to name just a few examples. It can be said that most of the contributions focus on issues different from the one described in this contribution and choose a different approach.

Our understanding of quotations in this project is limited to similar strings of words, i.e. the quotations are very close to borrowings or repetition of verbatim or almost verbatim passages. Technically, it can be viewed as an n-gram that is being repeated in at least two works. These repeated n-grams create links that exhibit some hierarchy, e.g. on the chronological line. The only approach known to us that can be paralleled to ours is the one described in Kolak and Schilit (2008) for quotation mining within the Google Books corpus with algorithm searching for verbatim quotations only.

In a different context and without direct inspiration we developed an algorithm that is tolerant to a certain degree of lexical and morphological variation and word order variability. The reason for this tolerance is both the type of the Arabic language (flective morphology and free word order), but also the fact that the quotations in medieval Arabic literature tend not to be very strict. Despite of the fact that the matching is not so rigorous, we assume that the length of n-grams we use drastically decreases possibilities of random matches.

The frequency of such n-gram repetition in various literary works can point to several aspects, however, in this contribution we will limit ourselves to interpreting such links in a rather cautious and not too far-reaching manner, mainly as pointers to the fact that the writer of the book where the quotations appear was also a reader of the book from which the quotations stem and that he was to a certain degree influenced by it.

This does not necessarily mean that the lineage of quotations is complete in our picture, for we

Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @ EACL 2014, pages 17–24,
Gothenburg, Sweden, April 27, 2014. ©2014 Association for Computational Linguistics

have to admit that there could be some author — member of the lineage — who is not involved in our corpus. In our graph, however, edges point to the first instance of a given n-gram in our data.

2 The Data, Its Organization and Extraction

It is obvious that for the type of the task mentioned in the previous chapter, there is a need of an appropriate data set.

2.1 Historical Corpus of Arabic

All the data in this contribution come from a historical corpus of Arabic (CLAUDIA — Corpus LinguæArabicæUniversalis DIAchronicus). This corpus covers all the main phases of the Arabic writings, from the 7th century to mid 20th century C.E. It contains ca. 2 thousand works and ca. 420 million words. The individual works are present in their entirety, i.e. each file contains a full text of a given literary work, based on edited manuscripts. All the main registers (genres) that appeared in the history of Arabic literature are represented in the corpus.

All the texts in the corpus are raw, without additional annotation. The files contain only a basic annotation of parts to be excluded from analyses (introductions, editorial forewords, etc.). This is of importance for the algorithms development, as the ambiguity of a text written down in Arabic letters is rather high (cf. e.g. Beesley 2001, Buckwalter 2004 or Smrž 2007 passim). On the other hand, it is certainly clear that the ambiguity significantly decreases when the n-gram information (i.e. context) is introduced.

As such, the corpus can be viewed as a network-like representation of Arabic literature. Each work is assigned several attributes, such as authorship, position on the time line, genre characteristics, etc. As several of the attributes can be viewed from several angles, it should be made clear that the genre characteristics currently used correspond to rather traditional terms used in Arabic and Islamic studies. Currently, the attributes assigned to the individual works are based on extra-corpus information and all of them were assigned manually from standard sources.

A short remark on the character of Arabic literature is appropriate. One should bear in mind that the approach to literature as consisting only of belles-lettres is relatively new, and for Arabic literature can be applied at the soonest at the end of the 19th century. All the previous phases must be seen as containing very different genres, including science, philosophy, popular reading and poetry as well as a huge bulk of writings connected with Islam, thus representing rather the concept of "Schrifttum" as expressed in the canonical compendia on Arabic literature, such as Brockelmann (last edition 1996). This is also reflected in current contribution, as many of our examples are connected with Islamic literature covering all the aspects of the study of religion. This includes theology, Islamic law, history, evaluation of sources, tradition, etc. Further information can be found e.g. in Kuiper 2010.

2.2 The Grid and the Network

The construction of a grid from a corpus consists basically in defining some constitutive units that serve as nodes. There are several possibilities of constituting such units, but some obvious solutions might not work very well. At first glance, it is advisable to find as small a unit as possible, while still retaining its meaningfulness; we decided to identify such units with individual works, or titles, with possible further division: Arabic literature is full of several-volume sets, and as our analyses showed, it may be sometimes useful to treat them as multi-part units, where individual parts can be treated as individual nodes (e.g., in some of our analyses it appeared that only a second volume of a three-volume set was significant). Treating such parts as individual nodes reveals similar cases instantly and can prevent overlooking important features during the analysis.

The nodes should allow reasonable links leading from one node to another. These links are crucial for any possible interpretation, as they show various types of relations between individual nodes. These nodes can be again grouped together, to show relations among different types of grouped information (i.e. links between titles or their parts, among authors, centuries, genres, etc.).

The nodes as such create the basis for the construction of both the grid and the network. As pointed out, currently the main axes used for grid and network construction are the authorship, chronological line, and the register information. The links among individual nodes are interpreted as relational links, or edges, in a network. These links also reflect quantitative data (currently, the number of quotations normalized to the lengths of the documents). The grid currently consists of the chronological line and the line of the works (documents). Above this grid, a network consisting of edges connecting the works is constructed. The grid in our approach corresponds to a firm frame where some basic attributes are used. The network then consists of the relations that go across the grid and reveal new connections between individual units.

A terminological remark is appropriate here. The network constructed above the grid corresponds to a great deal to what is called a *weighted graph* (the width of edges reflects the frequency of links). The term *directed graph* could also be used, however, in our current conception of the network, the links are not really oriented, as the direction of links pointing to contemporary authors is sometimes not clearly determinable, contrary to authors with greater time gap.[1] That is why we call these links *edges* and not *arcs*, and possibly, the graph could be called a *semi-directed graph*.

Kolak and Schilit (2008) observe that the standard plagiarism detection algorithms are useless for unmarked quotation mining and suggest straightforward and efficient algorithm for repeated passage extraction. The algorithm is suitable for modern English texts, since quotations are more or less verbatim and the word order is stable. But it is insufficient for medieval Arabic texts as the quotations are usually not really strict and the word order in Arabic is variable. We decided that our algorithm must be a) word order insensitive; b) tolerant to certain degree of variability in the content of quotations, so that the algorithm allows some variation introduced by the copyist, and reflects possibilities of change due to the fact that Arabic is a flective language.

2.3 Quotations extraction: technical description

The basic operation in the process is the quotations extraction. The procedure itself could be used in plagiarism detection, however, such labels do not make sense in case of medieval literature with different scales of values.

The quotation extraction process consists of four phases:

1. The corpus is prepared for analysis. Numerals and junk characters are removed from the corpus, as well as all other types of noise. Reverse index of all word types in the corpus is constructed (in case of texts written in Arabic script, a special treatment of diacritical signs and the *aliph*-grapheme and its variants is necessary).

2. All repeating n-grams greater than 7 tokens are logged (the algorithm is tolerant to the word order variability and to the variability of types up to 10 %) [2] : Tokens of every n-gram in the text are sorted according to their frequency in the whole corpus (for every n in some reasonable range, in our case $n \in < 7; 200 >$).

   (a) The positions of $round(0.1n) + 1$ least frequent tokens[3] are looked up in the reverse index.

   (b) The neighbourhoods of the positions are tested for being quotations of the length of *n* tokens.

   (c) Quotations are merged so that quotations larger than *n* tokens are detected as well.

3. For each pair of texts i, j the following index $\Xi_{(i,j)}$ is calculated (N is the number of tokens in a text, M is the number of tokens that are part of quotations of the text j in the text i, K is the set of all pairs of texts in the corpus; h is the parameter that determines number of edges visible in the graph, for details see below):

[1] Our time reference is based on the date of death of respective authors, and thus can be considered as "raw". Data on the publication of a respective book are often not available for more distant periods.

[2] The minimal length of the quotation and the percentage of word types variability should have been determined on an empirical basis, maximizing recall and precision. The problem is that the decision whether the repeating text passage is a quotation or not is not a binary one. Kolak and Schilit (2008) note the problem and let their subjects evaluate results of their algorithm on a 1–5 scale. As we did not manage to do vast and proper evaluation of the outputs of our algorithm using various minimal lengths of the quotations and degrees of variability, we relied on our subjective experience. The minimal length was set so that it exceeds length of the most common Arabic phrases and Islamic eulogies and the percentage of variable words was set to cover some famous examples of formulaicity in Arabic literature

It needs to be said that some minor changes of the parameters do not influence the results excessively, at least for the case studies we present here.

[3] The reason being the 10% tolerance.

$$\Xi_{i,j} = log_2 \frac{h\frac{M_{i,j}}{N_i N_j}}{\sum_{(k,l)\in K} \frac{M_{k,l}}{N_k N_l}}$$

It should be noted that the formula given above is inspired by the Mutual Information but it has no interpretation within the theory of information. It was constructed only to transform the number of quoted tokens into some index that could be graphically represented in some reasonable way convenient to the human mind.

4. The edges representing links with Ξ lower than a certain threshold are omitted. The threshold is set to 0.5 according to the limits of the programs producing graphic representation of the graph (the width of the line representing the edge is associated directly with the index Ξ). The index is normalized by the parameter h so that the user can set density of the graph, i.e. manipulate the index on an ad hoc basis with consideration for suitable number of edges and their ideal average width. E.g., the number of word tokens involved in autoquotations in Qur'an is 13 956 and the overall number of tokens is 80 047.

$$\frac{M_{Qur'an,Qur'an}}{N_{Qur'an}N_{Qur'an}} = \frac{13\,956}{80\,047^2} = 0.00000218$$

For our corpus, the average value is 0.000025, setting $h < 16.23$ then means that the Qur'anic autoquotation link will not be represented in the graph. Setting $h = 0.346574$ means that an average link gets $\Xi = 0.5$. Setting $h = 2$ means that an average link gets $\Xi = 1$.

The relation is exported to the .dot format and the graph is generated by popular applications GraphViz and GVEdit.[4]

The resulting database is stored in a binary format, but the graphical user interface allows the researchers to export graphs in accordance to their concepts. The features of the graphs can be changed by manipulating the h parameter and some other options. The appearance of the nodes can be freely adjusted as well.

More detailed information on the overall technical process is available directly from the authors.

3 The Analysis and Interpretation

The results are currently stored in a database and are available for further analyses. It is clear that results from a corpus of 420 million words offer many ways of interpretation.

The usage of the extracted data is to a certain degree limited in nature. It is mainly suitable for discussion of relations among individual nodes (documents, titles) or their groups. However, further processing of the data will enable a wider palette of possibilities. Currently, and also due to the limitations of this paper, only a few examples will be given.

3.1 Central Nodes and Relevance

The centrality of a given document may point to its relevance for its surroundings. If the relations that were found by our algorithms are interpreted merely as showing influence of predecessors on the author and his influence on his successors, then the number of links to and from an author and his particular book shows the relevance of that book.

In graph theory, there is no general agreement on how centrality should be defined. We expand the large number of indices of the degree centrality with our own index that is based on the same idea as the Ξ index (J is the set of all texts):

$$C_D(i) = \sum_{j\in J} \frac{M_{i,j}}{N_i N_j}$$

The measurement of this rather primitive and straightforward index results in table 1. The table also contains the plain number of edges at $h = 10$ (marked as *edg.*):

As the pointers to the subject of the respective works show, it was not only Islamic subjects that found their way to the most cited works in Arabic literature — historical literature as well as educative literature obviously played an important role in the medieval Arabic civilization.

It is interesting that az-Zayla'i's node comprises only the second volume of his three-volume *Nasab ar-Raya* (*Erection of the Flag*) — the other volumes exhibit either no edges or very few (0–1 and 1–0 respectively and the quotations point to his 2nd volume). Another interesting fact is that az-Zayla'i is rather less-known today — a short reference can be found in Lane 2005: 150 (fn. 2 and 3). This is also confirmed by the situation today. An Internet search for this author (including Arabic sources) yields only a short paragraph on his

[4] http://www.graphviz.org

	Degree C_D	Cited C_D	Citing C_D	Cited edg.	Citing edg.
1	0.0958	0.0278	0.0681	70	12
2	0.08257	0.0789	0.0036	23	5
3	0.07763	0.0001	0.0775	0	2
4	0.07277	0.0597	0.0130	155	0
5	0.04562	0.0038	0.0418	35	13

Table 1: Texts sorted according to the degree centrality (first five texts). Authors with their works and genre:
1 = az-Zayla‘i — *Nasab ar-Raya*, vol. 2 (Islam)
2 = Abu Nu‘aym al-Isbahani — *Axbar Isbahan* (history)
3 = Abu Nu‘aym al-Isbahani — *Tarix Isbahan* (history)
4 = an-Nasa’i — *Sunna* (Islam)
5 = al-Yafi’i — *Mir’at al-Jinan* (educative literature — adab).

birth (small village in Somalia, no date) and death (Cairo 1360).

Ibn Xaldun (d. 1382) is a very well-known figure today, respected for his *History*. Today, especially his *Introduction* (*Muqaddima*) is appreciated as an insightful methodological propedeutics. In Figure 2, his relevance in the Middle Ages is measured: it comprises 4 volumes: *Introduction* and *History* vols. 1–3. The graph shows (apart from numerous autoquotations) that his 3rd volume is the central one, where most of incoming and outgoing links can be found. On the other hand, his Muqaddima, which is praised today for its originality, remains isolated (our data do not cover the second half of the 20th century, where this appreciation could be found).

3.2 Time Depth

As our network combines a grid with chronological axis, it is rather easy to follow the distribution of links connected to a given node not only the relevance to other nodes, but also in time. As relevance of a given work is mostly judged from our current point of view (i.e. from what is considered important in the 21st century), an unbiased analysis may give interesting results showing both inspirational sources of a given work and its influence on other authors; it can also show the limits of such influence.

Figure 1 concentrates on the figure of az-Zayla‘i (d. 1360), who obviously played an important role in transmitting the knowledge (or discussion, at least) between different periods (cf. 3.1). The second volume of his *Nasab ar-Raya* is a clear center of the network.

The dating of the numerous sources that he used while writing his book starts ca. from the 10th century and to a great deal almost ignores 11th and 12th centuries. There is a thick web of links to his contemporaries, and his direct influence is very strong on the authors of the following century, but slowly wanes with the passage of time — although there are some attestations of his influence in the 16th and 17th centuries, they are getting less and less numerous. In the 20th century there are only two authors at whom we found some reflection of az-Zayla‘i ’s work.

From the point of view of the 21st century, az-Zayla‘i is a marginal figure, both for the Western and Arabic civilizations. On the other hand, as our data show, his importance was crucial for the discussion on Islamic themes for several centuries, which is, apart from the data given above, confirmed also by frequent quotations of his name and writings in the titles starting from the 15th century on.[5]

It is appropriate to repeat here that such conclusions can be viewed as mere signals, as we cannot exclude that there is some title occurring in the quotations lineage but missing in our data.

It should also be stressed that these conclusions reflect only verbatim quotations and are not based on the contents of these works. In other words, the relations do not represent an immediate reflection of the spread of ideas of a given author but rather show the usage of a given work in various periods of the evolution of Arabic literature.

4 Future Work

It is clear that there are many ways in which we can continue in our project. In the near future, we plan to work on the following topics:

- experimenting with various lengths of the shortest quotation and the degree of allowed variability, maximizing recall and precision.

[5] The title of the book is attested in other writings in our dataset in the 15–17th centuries only; the name of the author appears abundantly in the 15th century (ca 1050x), 16th century (ca 560x), 17th century (ca 500x). The 18th century gives only 45 occurrences, later on his name can be found only in specialized Islamic treatises.

- enriching the palette of nodes' attributes to enable a broader scope of analyses based both on external sources and inner textual properties of given texts;

- comparison of the complexity of the graphs of various subcorpora organized according to different criteria;

- comparison of various indices of centrality;

- detailed interpretation of edges;

- comparison with other corpora and

- network of autoquotations within one text.

Acknowledgments

The research reflected in this article has been supported by the GAČR (Czech Science Foundation), project no. 13-28220S.

We would also like to thank to the anonymous reviewers for their inspiring comments.

References

Kenneth R. Beesley. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. *ACL Workshop on Arabic Language Processing: Status and Perspective.* Toulouse, France: 1–8.

Carl Brockelmann. 1996. *Geschichte der Arabischen Literatur,* (4 Volume Set). Brill, Leiden (1st edition: 1923).

Tim Buckwalter. 2004. Issues in Arabic Orthography and Morphology Analysis. *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING.* Geneva: 31–34.

William Paulo Ducca Fernandes, Eduardo Motta and Ruy Luiz Milidiú. 2011. Quotation Extraction for Portuguese. *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology.* Cuiabá: 204–208.

Okan Kolak and Bill N. Schilit. 2008. Generating Links by Mining Quotations. *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia.* New York: 117–126.

Kathleen Kuiper. 2010. *Islamic Art, Literature and Culture.* Rosen Publishing Group.

Andrew J. Lane. 2005. *A Traditional Mu'tazilite Qur'an Commentary: The Kashshaf of Jar Allah al-Zamakhsari (d.538/1144).* Brill, Leiden.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Seattle: 989–999.

Bruno Pouliquen, Ralf Steinberger and Clive Best. 2007. Automatic Detection Of Quotations in Multilingual News. *Proceedings of Recent Advances in Natural Language Processing 2007.* Borovets.

Xiaolin Shi, Jure Leskovec and Daniel A. McFarland. 2010. Citing for High Impact. *Proceedings of the 10th annual joint conference on Digital libraries.* New York: 49–58.

Otakar Smrž. 2007. *Functional Arabic Morphology. Formal System and Implementation.* Doctoral Thesis, Charles University, Prague.

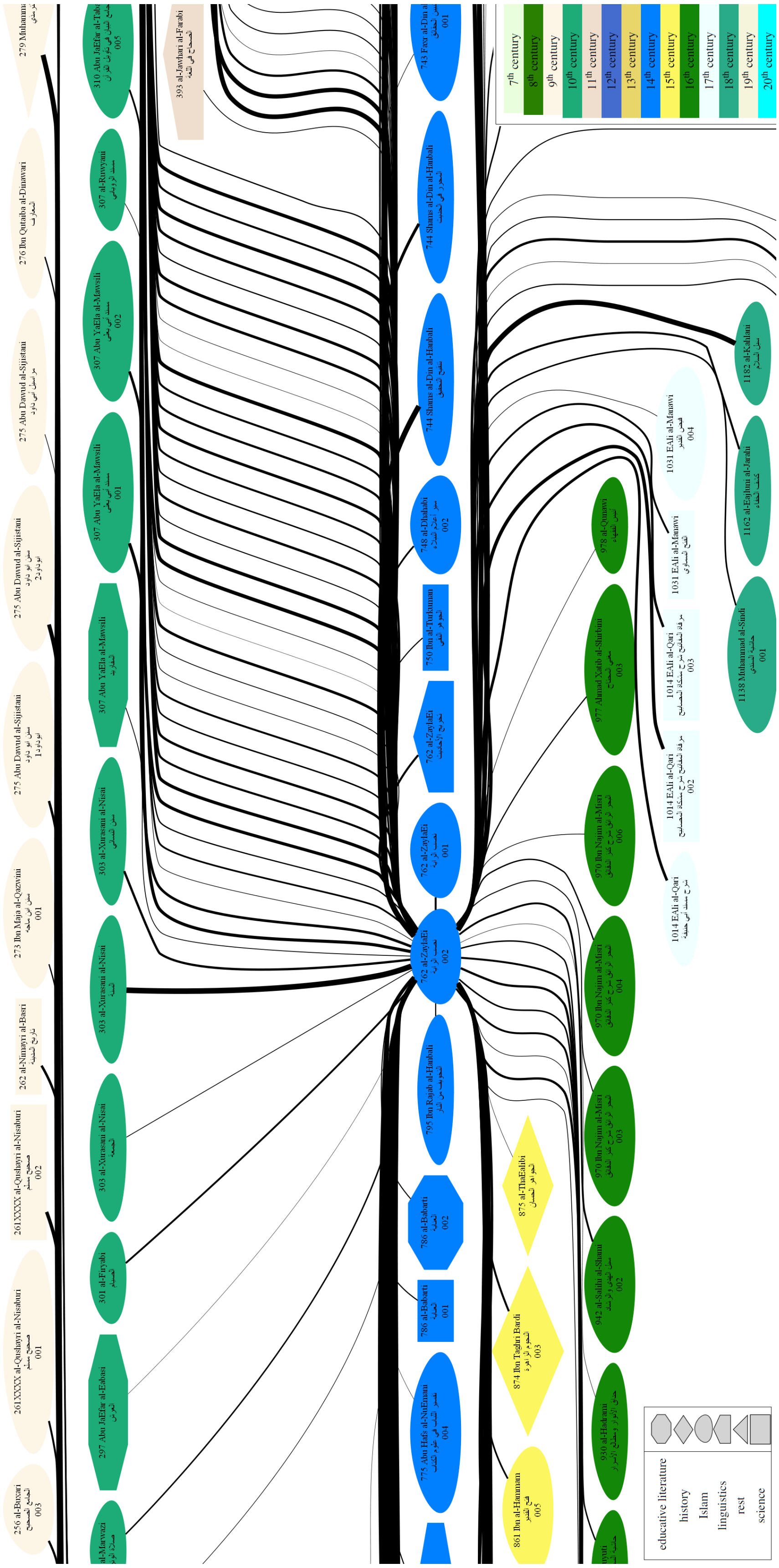

Figure 1: Case study: Zayla‘i’s Nasab ar-Raya 3 in its context. Parameter $h = 2$. Cut out.

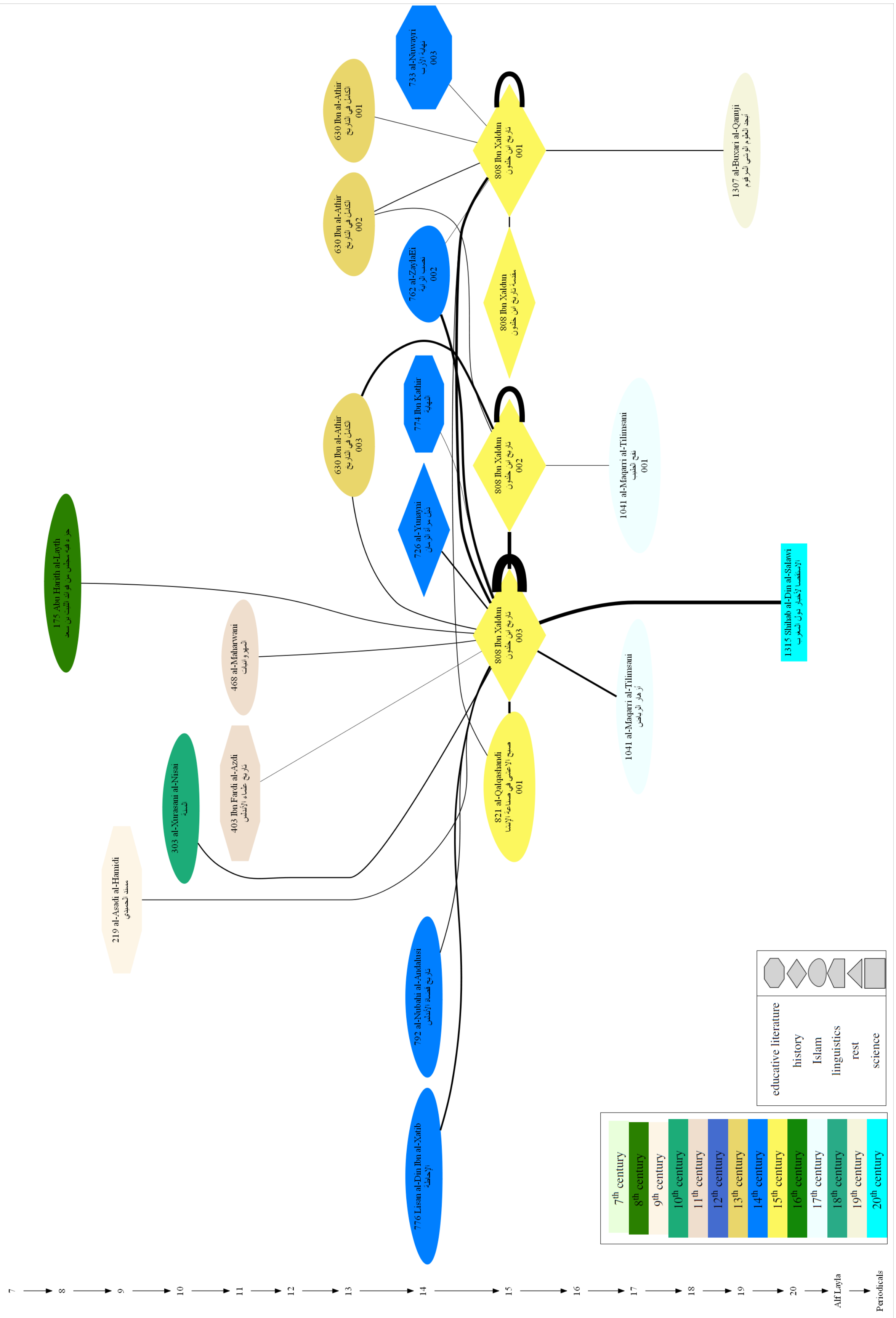

Figure 2: Case study: the network around the Ibn Xaldun's works. Parameter $h = 1.6667$.