Instituto Federal do Rio Grande do Norte
Campus Natal - Central
Diretoria de Gestão e Tecnologia da Informação
Tecnologia em Análise e Desenvolvimento de Sistemas

Exploring IDRs (Interesting Dense Regions) and Outliers in a Spatial-Temporal Data Environment

Francisco Bento da Silva Júnior

Natal-RN
Dezembro, 2018

Francisco Bento da Silva Júnior

Exploring IDRs (Interesting Dense Regions) and Outliers in a Spatial-Temporal Data Environment

Trabalho de conclusão de curso de graduação do curso de Tecnologia e Análise em Desenvolvimento de Sistemas da Diretoria de Gestão e Tecnologia de Informação do Instituto Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Tecnologo em Análise e Desenvolvimento de Sistemas.

Linha de pesquisa:
Nome da linha de pesquisa

Orientador

Nome completo do orientador e titulação

TADS – Curso de Tecnologia em Análise e Desenvolvimento de Sistemas
DIATINF – Diretoria Acadêmica de gestão e Tecnologia da Informação
CNAT – Campus Natal - Central
IFRN – Instituto Federal do Rio Grande do Norte

Natal-RN

Dezembro, 2018

Trabalho de Conclusão de Curso de Graduação sob o título *Título* apresentada por Nome completo do autor e aceita pelo Diretoria de Gestão e Tecnologia da Informação do Instituto Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Nome completo do orientador e titulação
Presidente
DIATINF – Diretoria Acadêmica de Gestão e Tecnologia da Informação
IFRN – Instituto Federal do Rio Grande do Norte

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Instituto

Nome completo do examinador e titulação
Examinador
Diretoria/Departamento
Universidade

Natal-RN, data da defesa (dia, mês e ano).

Homenagem que o autor presta a uma ou mais pessoas.

Agradecimentos

Agradecimentos dirigidos àqueles que contribuíram de maneira relevante à elaboração do trabalho, sejam eles pessoas ou mesmo organizações.

Citação

Autor

Exploring IDRs (Interesting Dense Regions) and Outliers in a Spatial-Temporal Data Environment

Autor: Francisco Bento da Silva Júnior
Orientador(a): Titulação e nome do(a) orientador(a)

Resumo

O resumo deve apresentar de forma concisa os pontos relevantes de um texto, fornecendo uma visão rápida e clara do conteúdo e das conclusões do trabalho. O texto, redigido na forma impessoal do verbo, é constituído de uma sequência de frases concisas e objetivas e não de uma simples enumeração de tópicos, não ultrapassando 500 palavras, seguido, logo abaixo, das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores. Por fim, deve-se evitar, na redação do resumo, o uso de parágrafos (em geral resumos são escritos em parágrafo único), bem como de fórmulas, diagramas e símbolos, optando-se, quando necessário, pela transcrição na forma extensa, além de não incluir citações bibliográficas.

Palavras-chave: Palavra-chave 1, Palavra-chave 2, Palavra-chave 3.

Exploring IDRs (Interesting Dense Regions) and Outliers in a Spatial-Temporal Data Environment

Author: Francisco Bento da Silva Júnior
Supervisor: Titulação e nome do(a) orientador(a)

Abstract

O resumo em língua estrangeira (em inglês *Abstract*, em espanhol *Resumen*, em francês *Résumé*) é uma versão do resumo escrito na língua vernícula para idioma de divulgação internacional. Ele deve apresentar as mesmas características do anterior (incluindo as mesmas palavras, isto é, seu conteúdo não deve diferir do resumo anterior), bem como ser seguido das palavras representativas do conteúdo do trabalho, isto é, palavras-chave e/ou descritores, na língua estrangeira. Embora a especificação abaixo considere o inglês como língua estrangeira (o mais comum), não fica impedido a adoção de outras linguas (a exemplo de espanhol ou francês) para redação do resumo em língua estrangeira.

Keywords: Keyword 1, Keyword 2, Keyword 3.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

Sumário

1 Introduction

In the last ten years, the search for terms such as big data, data analysis, and data visualization has increased enormously. One of the reasons is that with the advancement of technology and computers, we have been able to generate huge masses of data from different sources in various formats and in an incredibly small time. Along with this came also new difficulties in the field of data analysis which is: How to process these immense quantities quickly and efficiently? How to visualize this amount of data? How to clean the dataset without losing important points?

When it comes to the data analysis field and the analyst is working with large datasets, is very common to have a lot of points with attributes very distant from the rest of the dataset. This happens because when more huge is your dataset, more easily you can find abnormal points that will be more distant from the normal distribution. This kind of behavior is important to the analyst study and discover more information about the dataset itself and with this, he could take more accurate decisions and propose better statements. Usually, this concern about the around data is not too relevant for the most researches, but recently it has appeared more frequently researches focused on this kind of data. These specific data with those characteristics is called *Outliers* and is very important that the current data analysts pay more attention to those data, because important information may be hidden inside these abnormal data. For example, if we get an isolated dataset about NYC taxis from 20XX and analyze the frequency of requests at the entire year, it will appear a lot of points very distant from the average curve and this will indicate an unusual behavior in this collection. Usually, the first step to take in this situation is to remove the irregular points and continue the processing with the rest of the dataset, but if we get another isolated dataset about the wind speed from the same NYC region and compare this exactly space of time, we will perceive few peaks of high speed indicating hurricane at the same time (FREIRE et al., 2016). Analyses like this prove the importance of detect, study and interpret those outliers to increase the knowledge obtained from this dataset.

1.1 Context

Nowadays we are more and more connected with multiple applications that access a huge amount of our existing data and even generate more to improve their analyses about us for diverses proposals. Tools like Google Maps, Uber, Waze has a lot of realtime spatial data about our traffic behavior (private cars, public transportation, taxis, etc.), work place, travel location, etc.

When it comes to regular users, it is very common that he will be lost in such masses of spatial data and this will damage your possible analyse, even the simplest one. This commom problem still does not have an definitively solution, so existing researches trying to indicate possible approaches for mitigate this problem and be close to a working solution. These approaches are based on: agroup a large amount of data by specific attributes and summarize the commom attributes between those data for give simple insights about these groups, filter the dataset to reduce the showing possibilities and focus on specific data for a more detailed (but not wide) analysis, and a lot of others strategies to reduce the complexity of the analysis.

Together with those most commom problems, exists an important one that can happen before the first step of analysis that is: What do when parts of the dataset seems to be irregular or with corrupted data? There are techniques that helps to clean those parts and not compromise the analysis, but recently studies points the importance of these *"abnormal"* data and how much the analyst can learn just studying more precisely this set. (FREIRE et al., 2016)

In this complex environment of spatial data analyses with a lot of variables and possibilities, an user can easily fail in some of these step compromising severely the results of his analysis. Combining all these details, we suggest an approach that take as relevant the user feedback (capturing the mouse track) and based on those feedback, we will be able to analyse the user interest and inside this we can detect, study and propose actions to perform when an data considered outlier appears in this region of user interest.

Aiming at this problem, several types of research and tools have appeared trying to solve or improve it in some way, either by proposing techniques to increase the performance of the analyzes or to perform the data cleaning or to improve the structure of how to save these data. Among these researches there is a part focused on how to visualize these large amounts of data and even more when it comes to spatial data, as it turns out to be a

serious problem the more the amount of data grows, since the researcher could end up getting "lost "in the middle of so much information leaving their analysis greatly damaged.

In this context, one of these researches produced a new proposal that aims to improve the visualization and analysis of huge amounts of spatial data, the GeoGuide: a tool in which it is possible to load a generic dataset with spatial data (latitude and longitude attributes) and metadata and then visualize it on a global map to better navigate between them. Along with this there is also the concept of diversity and similarity that serves for an approach in which the researcher expands its area of research through highlights of similar points in distinct areas of a single point chosen by him. for example: Joana is a culinary enthusiast and wants to find new restaurants in neighborhoods that serve Brazilian food at a price range from $20 to $100. In a few clicks, you can find these suggestions in GeoGuide.

However, a new problem arises that is the availability of Joana to be able to reach certain neighborhoods because she does not have a car and needs public transportation to transit in her city. Aiming at this new feature, GeoGuide is adding the concepts of regions of interest (neighborhoods, in the case of Joana) so that the researcher can, implicitly (using the mouse movement), demonstrate which region is more interesting for him and thus avoid one more step that would be the process to exclude the suggestions of the GeoGuide that would be in unavailable places for Joana to access. With this concept it is also possible to solve another problem that would be the case of Marcos, who is passionate about travel and wants to redo a trip to Italy, however, he decided that he does not want to visit the same sights. So to avoid a new process, Mark would get suggestions outside his region of interest (which would be his last places visited in Italy), and then enjoy his trip.

1.2 Objectives

In this section are defined the general and specific objectives of the work.

1.2.1 General Objectives

- Introduce the problem of data analysis and visualization in large spatiotemporal datasets nowadays.
- Explain our proposed approach to detect spatial outliers in large datasets using the

concept of IDR and capturing user's feedback.

- Present our results using the given approach to detect outliers in our spatial-temporal environment and the benefits of these experiments.

1.2.2 Specific Objectives

- Analyze the latest researches in the field of outlier detection in spatial-temporal datasets.
- Present our proposed tool for spatial-temporal data analysis and visualization.
- Compare the presented researches showing the pros and cons of each work.
- Describe the concept of IDR used in our tool to mapping the user preference in a spatial-temporal environment.
- Summarize the most known existing outlier detection algorithms for generic and spatial data.
- Display our chosen outlier detection algorithm and explain the reasons for this choice.
- Apply our IDR concept and our chosen outlier detection algorithm in a spatial-temporal data environment.
- Present the results of our application and indicate our future work.

1.3 Methodology

TODO: Methodology

1.4 Work Organization

The rest of the document is organized as follows. Section 2 summaries the existing researches in data analysis and visualization field comparing with our proposed tool. Section 3 describes the concepts of IDRs (Interesting Dense Region) and Outliers with the existing algorithms for their detection and our chosen algorithm to detect outliers in our platform. Section 4 explains how we apply the IDRs and outliers detection in the

GeoGuide tool. Section 5 presents two applications using distinct datasets to demonstrate the applicability of it to real-world problems and the advantages of this approach. It shows and discusses the outlier detection results. Finally, a conclusion and some directions for future works are given in Section 6.

2 Background

In this chapter will be presented the existing researches about data analysis and outlier detection with algorithms and strategies of how to use this approach to improve the analysis of your data and increase the amount and the quality of the information that you can retrieve from your datasets.

2.1 Related Work

2.1.1 GeoGuide

Pursuing improve the spatial data analysis and the guidance approach for this kind of data, the GeoGuide (OMIDVAR-TEHRANI et al., 2017) is a interactive framework that aims to highlight to the analyst a subset of k interesting spatial points, based on the analyst *implicit* (e.g. mouse tracking) and *explicit* (e.g. points clicked) feedbacks, that he may not seem because of the huge amount of information on his screen. This framework considers two metrics to give the subset highlight. The first one is the **relevance** of each point to the point selected by the analyst considering the attributes of those points. The second one is the geographically **diversity** to expand the analyst area in chase of possible new interesting regions. All this process can be used in generic spatial datasets, as long as each point have two characteristics: geographical attributes (i.e. latitude and longitude) and metadata attributes about its own domain. For instance, the Airbnb platform [1] has open datasets about the available home-stays to rent and each one has the geographical attributes and *price, hostname, availability* as their metadata attributes that are specific for each dataset type. Using this approach, the GeoGuide is the first efficient interactive highlighting framework for spatial data, combining analyst feedbacks with relevance and diversity metrics to display to the analyst a set of interesting points that he may be not focused during his analysis.

[1] *http://www.airbnb.com*

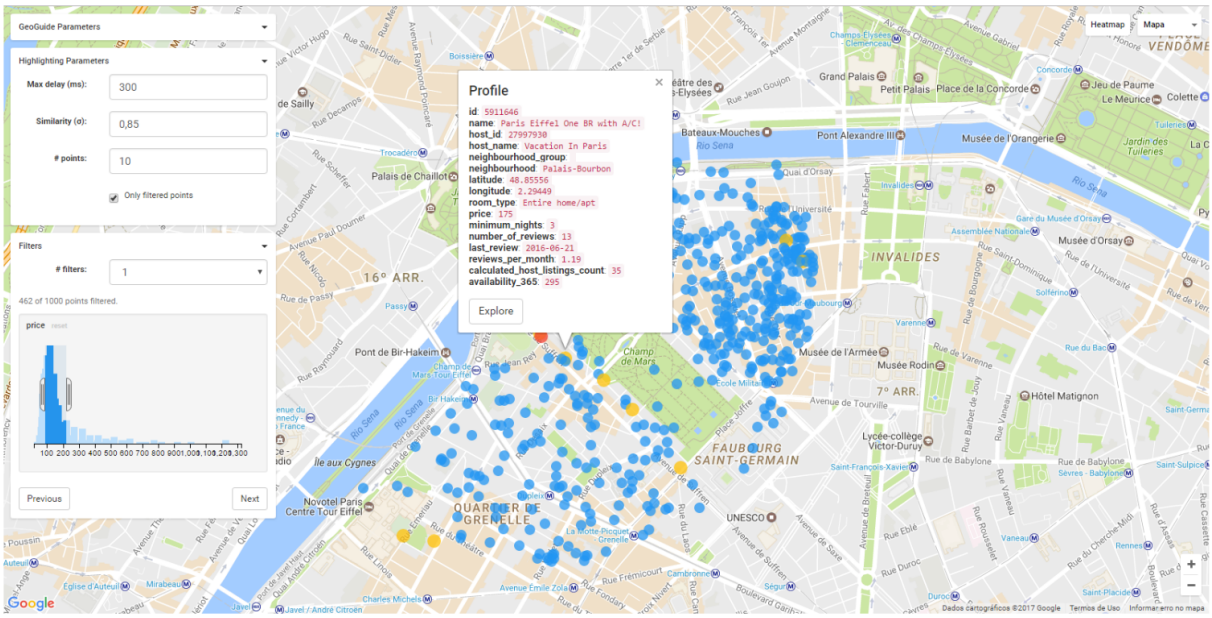

Figura 1: GeoGuide Image on Airbnb Dataset - Paris City

2.1.2 Outliers

Outliers in the statistics area are when one finds "aberrant values" in a given series of data, that is when one finds an atypical value or with a great distance from the normal distribution in that set. For example, when a researcher wants to monitor the temperature of his CPU during a certain time interval and it has been realized that the temperature range is between 34 °C and 45 °C degrees being 48 °C the maximum temperature and 27 °C the minimum temperature and in the middle of this sample are some punctual registers of 0 °C, this can be characterized as an outlier and, most likely, will be understood as a malfunction of the equipment that performed the collection of these CPU temperatures.

However, there are several ways to interpret an Outlier (not only as a collection error), but also as: data that belong to a different population of the sample, a damaged data, areas in which a certain theory is not valid or even, when the sample is too large, it is normal to have some small amounts of outliers in that group. In cases where it is proven that it is not the fault of a collection equipment malfunction or that it was not a human mistake, it is extremely important to know the why of that outlier and try to understand it, because it is not interesting for a research simply remove it from the sample or re-signify it by assigning a new value. This change may compromise the validity of the research, and if this is done, it is extremely important to document and record those changes.

As the information technologies improve and continuously increase their computational power, a great variety of algorithms for outlier detection has been surging and applied

in different contexts with diverse characteristics and the choice for one of those algorithms is based on the problem domain. Next section will present some of those algorithms with a brief explanation about each one.

2.2 Algorithms

2.2.1 Z-Score

Z-Score is one of the simplest methods for detecting outliers in a dataset. It is a parametric method and takes into account only one attribute per execution. It is also necessary the input of a threshold (usually is a value between 2.5 to 3.5) to be able to define if a given data can be considered an outlier or not. This method is suitable for small datasets that follow the Gaussian distribution.

2.2.2 DBSCAN

It is a density-based spatial clustering algorithm (ESTER et al., 1996) that can be applied in datasets that cannot be presumed what their distribution. It accepts multidimensional datasets (with 3 or more dimensions). However, you need a parameter (MinPts) that defines how many minimum points are needed to form a cluster. Thus, if the size of the set change, this parameter will need to be updated, otherwise the DBSCAN can become inefficient.

2.2.3 Isolation Forests

It is an algorithm of detection of outliers (LIU; TING; ZHOU, 2008) that uses a concept of machine learning that is the decision tree. It is one-dimensional (only takes one attribute at a time) and is required few parameters (this facilitates the configuration and use of the algorithm). No need to climb your values and a very robust algorithm for large datasets.

2.2.4 FDC

It is a depth-based algorithm (JOHNSON; KWOK; NG, 1998) approach for detection of outliers in 2D datasets based on the concept of ISODEPTH algorithm. The FDC computes the first k 2D depth contours (the points that can be considered outliers) by restricting to a small part of the complete dataset. In this way, it is more efficient by not having to

calculate in the complete dataset and thus scaling more easily for large datasets of two dimensions.

2.2.5 HOD

It is a distance-based outlier detection method (XU et al., 2016) that emerges to overcome the statistics-based concept because in the vast majority of datasets the probability distribution is not known. In this way, the method search for outliers based on their distance from their neighbors and if that point has a distance greater than a predefined parameter, then that point is considered an outlier. However, if there is a cluster of outliers in the dataset, this can affect its detection by distance-based algorithms, with this comes the concept of HOD (Hidden Outlier Detection) algorithms that aim to find outliers even when they are grouped and in enough quantity to form a cluster.

2.2.6 ORCA

It is a distance-based algorithm (BAY; SCHWABACHER, 2003) that optimizes a simple nested loop algorithm (which are logarithmic algorithms and extremely inefficient when dealing with large datasets) by removing possible non-outliers during their execution. This way, instead of processing the complete dataset by calculating all possible distances, it removes unnecessary calculations that would be executed if a non-outlier point were taken to the end. From him, new researches have emerged further refining this concept.

2.2.7 Linearization

It is a distance-based algorithm (ANGIULLI; PIZZUTI, 2002) that detects outliers by calculating the sum of the distances of a point in relation to its neighbor, calling it weight, and setting an outlier as the points with the greatest weights in the dataset. In this way, it is an efficient algorithm and it is linearly scaled both in the number of points and in the number of dimensions. To calculate these outliers more efficiently the algorithm uses the concept of the Hilbert space-filling curve.

2.2.8 RBRP

It is an algorithm for high-performance multidimensional datasets that is based on distances between the points to be able to define what the outliers are (GHOTING; PARTHA-

SARATHY; OTEY, 2006). Its difference to the other distance-based algorithms is that it is more efficient for datasets with multiple dimensions and in comparisons with others, its scalability is approximately linear for the number of dimensions and logarithmic for the number of points in the dataset.

2.2.9 LOF

It is a density-based algorithm that adds a new concept in the search for outliers: the Local Outlier Factor (LOF) (BREUNIG et al., 2000), which is a degree of propensity to be an outlier so that the process of outlier definition is not more binary, but something gradual. With this, the approach is not to define whether a point is an outlier or not, but rather the "how outlier"that point is in that dataset. The outlier factor is local in the sense that only a neighborhood of that point is taken into account to define its factor.

2.2.10 ABOD

It is an angle-based algorithm (KRIEGEL; HUBERT; ZIMEK, 2008) for detection of outliers that is focused on high-dimensional datasets, different from other distance-based algorithms that end up damaged when one has many dimensions. Your approach is based on the calculation of a degree of angle between the different vectors of a point with its neighbors. With this, more centralized points within the cluster will have this degree calculated with a high value, the points more on the edge of the clusters will have this degree a little smaller and the possible outliers will have that degree with a very small value, since they will generally be far from the cluster in a particular direction.

3 Considerações finais

As considerações finais formam a parte final (fechamento) do texto, sendo dito de forma resumida (1) o que foi desenvolvido no presente trabalho e quais os resultados do mesmo, (2) o que se pôde concluir após o desenvolvimento bem como as principais contribuições do trabalho, e (3) perspectivas para o desenvolvimento de trabalhos futuros, como listado nos exemplos de seção abaixo. O texto referente às considerações finais do autor deve salientar a extensão e os resultados da contribuição do trabalho e os argumentos utilizados estar baseados em dados comprovados e fundamentados nos resultados e na discussão do texto, contendo deduções lógicas correspondentes aos objetivos do trabalho, propostos inicialmente.

3.1 Principais contribuições

Texto.

3.2 Limitações

Texto.

3.3 Trabalhos futuros

Texto.

Referências

ANGIULLI, F.; PIZZUTI, C. Fast outlier detection in high dimensional spaces. In: ELOMAA, T.; MANNILA, H.; TOIVONEN, H. (Ed.). *Principles of Data Mining and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 15–27. ISBN 978-3-540-45681-0.

BAY, S. D.; SCHWABACHER, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2003. (KDD '03), p. 29–38. ISBN 1-58113-737-0. Disponível em: <http://doi.acm.org/10.1145/956750.956758>.

BREUNIG, M. M. et al. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 29, n. 2, p. 93–104, maio 2000. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/335191.335388>.

ESTER, M. et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <http://dl.acm.org/citation.cfm?id=3001460.3001507>.

FREIRE, J. et al. Exploring what not to clean in urban data: A study using new york city taxi trips. *IEEE Data Eng. Bull.*, v. 39, n. 2, p. 63–77, 2016. Disponível em: <http://sites.computer.org/debull/A16june/p63.pdf>.

GHOTING, A.; PARTHASARATHY, S.; OTEY, M. E. Fast mining of distance-based outliers in high-dimensional datasets. In: *Proceedings of the 2006 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2006. p. 609–613. Disponível em: <https://doi.org/10.1137/1.9781611972764.70>.

JOHNSON, T.; KWOK, I.; NG, R. Fast computation of 2-dimensional depth contours. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1998. (KDD'98), p. 224–228. Disponível em: <http://dl.acm.org/citation.cfm?id=3000292.3000332>.

KRIEGEL, H.-P.; HUBERT, M. S.; ZIMEK, A. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. (KDD '08), p. 444–452. ISBN 978-1-60558-193-4. Disponível em: <http://doi.acm.org/10.1145/1401890.1401946>.

LIU, F. T.; TING, K. M.; ZHOU, Z. Isolation forest. In: *2008 Eighth IEEE International Conference on Data Mining*. [S.l.: s.n.], 2008. p. 413–422. ISSN 1550-4786.

OMIDVAR-TEHRANI, B. et al. Geoguide: An interactive guidance approach for spatial data. In: *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, United Kingdom, June 21-23, 2017*. [s.n.], 2017. p. 1112–1117. Disponível em: <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.170>.

XU, H. et al. Index based hidden outlier detection in metric space. *Scientific Programming*, Hindawi Limited, v. 2016, p. 1–14, 2016. Disponível em: <https://doi.org/10.1155/2016/8048246>.

APÊNDICE A – Primeiro apêndice

Os apêndices são textos ou documentos elaborados pelo autor, a fim de complementar sua argumentação, sem prejuízo da unidade nuclear do trabalho.

ANEXO A – Primeiro anexo

Os anexos são textos ou documentos não elaborado pelo autor, que servem de fundamentação, comprovação e ilustração.