公募企画
# 公募企画シンポジウム6
# 安心・安全なビッグデータの流通プラットホームとセキュリティ基盤技術
2017年11月22日(水) 08:45 〜 10:45 C会場 (10F 会議室1001)

[3-C-1-PS6-4] 多機関分散データ統合活用技術による生活安全：学校事故データを用いた検証

西田 佳史, 北村 光司 （国立研究開発法人 産業技術総合研究所）

子ども、高齢者、障がい者などの生活機能変化者に対する生活安全が強く求められている。本研究では、セキュリティ基盤技術を活用した多機関分散データの利活用技術を、データ保有機関である医療機関等やデータ活用者である製品デザイン等の現場と連携し実証的に開発することを目的としている。これまでに、消防庁、日本スポーツ振興センター、複数の医療機関、保育所・小中学校などと連携し、分散された傷害関連データを統合的に活用する技術の開発を進めている。

具体的な応用分野の一つとして、学校安全を取り上げ、実証を進めている。昨今の学校安全に対する関心の高まりから、複数の小学校に分散する事故データを、学校を特定されない方法で、仮想的に統合することでビッグデータ化し、これに基づいた統計解析を行うことで、起こりうる重傷事故を予測したり、対策が必要となる重症事故を見つけ出すニーズが高まっており、これに対応するためにセキュリティに配慮しながら、医療費などのKPI付きビッグデータを作成し、重症となる事故要因を分析する技術（重症度クリフ分析技術）を開発した。

クリフ分析の概要を述べる。学校において、安全担当の教師などが入力した条件に合致する傷害データを複数の小学校から取得し統合する。得られた傷害データの事故状況記述データを対象にテキストマイニングを行い、事故状況を表すのに重要な単語を抽出する。重要な単語を特徴量として、クラスタ分析を行い、事故状況を分類する。医療費を重症度の指標として、重症度の変化点を見つけ出し、類似事故状況における重症事故群と軽傷事故群に分類して、それぞれの特徴を把握する。

また、本 CREST（ビッグデータ統合利活用促進のためのセキュリティ基盤技術の体系化）で連携している大阪大学らが開発したプライバシを保護したデータ集合演算技術（ PSI）と統合することで、セキュリティに配慮したクリフ分析が利用可能なシステムを開発した。

多機関分散データ統合活用技術による生活安全

西田佳史[*1], 北村光司[*1]
*1 産業技術総合研究所, 科学技術振興機構 CREST

Living Safety by Technology for Integratively Utilizing Multi-organizational Distributed Big Data

Yoshifumi Nishida[*1], Koji Kitamura[*1]
*1 National Institute of Advanced Industrial Science and Technology, CREST, JST

Living safety technology is strongly needed for child, elderly, persons with disabilities. But it is difficult to understand everyday life related problem based on injury data, medical data, and so on. Because such kind of data are distributed in multi-organization and due to privacy protection it is difficult to share and integrate them. To solve this issue, our project is developing technologies for integratively utilizing multi-organizational distributed big data based on security technology.

The authors research on methods for applying developed technologies to school safety. In this paper, we propose a system for analyzing high risk injury by integrating data of injury occurred in each school while keeping secrets. By treating medical cost data as KPI, the system can analyze high risk injury and enables to understand factors related to the difference between high risk injury and low risk injury. We conducted verification of availability of the developed system by applying to actual injury data at schools.

Keywords: Multi-organizational Distributed Big Data, Injury Prevention, School Safety

1. 緒論

子ども, 高齢者, 障がい者などの生活機能変化者に対する生活安全が強く求められている. しかしながら, 生活安全を考える上で重要な過去に起きた事故データや医療データなどが, 多機関に分散して存在しており, 全体像を理解することが難しくなっている. それらのデータを統合することができれば, データにもとづいた生活安全上の理解が可能となり, 対策を取ることが可能となる. しかし, 機関をまたいだデータの共有や統合は, プライバシ保護や情報漏えいなどの問題で難しい. この問題に対して, JST CREST(ビッグデータ統合利活用促進のためのセキュリティ基盤技術の体系化)プロジェクトでは, セキュリティ基盤技術を活用した多機関分散データの利活用技術の開発を進めている. 著者らの研究グループは, 生活安全分野を対象に, データ保有機関である医療機関等やデータ活用者である製品デザイン等の現場と連携し実証的に, 多機関分散データの利活用技術を開発することを目的としている. これまでに, 消防庁, 日本スポーツ振興センター, 複数の医療機関, 保育所・小中学校などと連携し, 分散された傷害関連データを統合的に活用する技術の開発を進めている.

具体的な応用分野の一つとして, 学校安全を取り上げ, 実証を進めている. 昨今の学校安全に対する関心の高まりから, データにもとづいて, 起こりうる重傷事故を予測したり, 対策が必要となる重症事故を見つけ出すニーズが高まっている. そこで, 本研究の目的は, 複数の小学校に分散する事故データを, 学校を特定されない方法で, 仮想的に統合することで医療費などのKPI付きビッグデータを作成し, 重症となる事故要因を分析する技術を開発することである. 本稿では, 学校環境下での傷害予防における課題について述べ, それを解決するシステムを提案する. また, 実際の学校環境下での事故データに適用することで, 提案システムの有用性を検証する.

2. 学校環境下での傷害予防の課題

学校環境下の事故による傷害を予防するためには, 過去発生した事故のデータに基づいて重症事故が起こり得る状況を把握し, 対策を取ることが重要である. しかし, 個々の学校で事故が頻発することはなく, 特に重症事故の頻度はさらに低いため, 実際に発生するまで把握できないことが多く, 予防が難しい. 学校環境下は類似した要素で構成されており, そこでの活動も類似しているため, 他の学校環境で発生した事故の情報も十分に活用可能である. そのため, 複数の学校間で事故データを共有して潜在リスクを把握したり, 事故データの分析者が, 複数の学校のデータを統合して, 仮想的にビッグデータ化することで, 学校環境での課題を把握することが可能となる.

しかし, 学校環境で発生した事故の情報は, プライバシ情報を含んでいたり, どこの学校で発生した事故かは知られたくないといった事情から, データを共有する仕組みが確立されていない. そこで, どの学校で発生した事故かは秘匿したまま, 事故状況の把握や傷害予防に必要な情報のみ共有・統合可能な仕組みが必要である.

複数機関で事故情報を共有して活用する場合, データを単純に共有しても活用が難しい. そのため, 特にリスクが高い事故の情報を提示したり, 実際に発生した事故の状況と類似した状況で発生した重症度が高い事故の情報を提示する, といった仕組みも必要である.

3. プライバシに配慮した多機関の事故情報の統合による重症事故分析システム

3.1 PSI を用いたプライバシを保護した情報共有

事故データを活用して学校環境下での傷害予防を行うためには, プライバシを保護した状態で, 複数機関でデータを共有したり, 統合したりして, 活用可能な仕組みが必要である. 著者らは, JST CREST(ビッグデータ統合利活用促進のためのセキュリティ基盤技術の体系化)のプロジェクトを実施して

おり，その中で開発されたプライバシを保護したデータ集合演算技術（PSI）[1]を用いた事故データの共有・統合システムを提案する（図1）．PSI は，指定したデータ項目に関して，複数機関の間でデータを暗号化したまま，共通集合を取り出すことが可能な技術である．この PSI を学校で発生した事故のデータを対象にして実行可能な仕組みにすることで，ユーザが取得したい情報の条件を指定することで，その条件に合致する事故情報を，いずれの機関もどの学校で起きた事故かは知らずに，統合した状態でユーザへと提供が可能になる．ここでの“いずれの機関も”とは，各学校はもちろんのこと，事故データを置いているサーバ，ユーザと学校との間に入るサービスプロバイダもどこの学校で起きた事故かは把握できない．このような仕組みにすることで，第三者である事故データの分析者だけでなく，現場の学校も自身の学校で起きた事故データをこの仕組みを使って共有することで，自身の学校で起きた事故とは知られずに，他の学校で起きた類似事故を把握する，といった活用が可能となる．

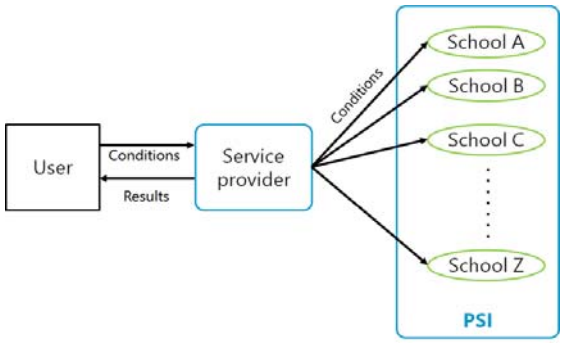

図1 PSI を用いた事故データ共有・統合システムの構成図

3. 2 重症度クリフ分析技術

PSI を用いた情報共有の仕組みを使うことで，複数の学校の事故データを統合して取得可能となるが，そのデータから有用な情報を取得するには手間が掛かる．そこで，医療費を重症度と捉えて，重症となる事故要因を分析する，重症度クリフ分析技術を開発した．この手法は，類似した状況の事故を対象に，重症度を分析し，特に重症度が高くなる変化点を見つけ出すことで，重症事例を把握することが可能である．また，重症度の変化点を境に，重症事故と軽症事故に分けて，その違いを分析することで，重症事故の要因を分析することが可能である．以下に，重症度クリフ分析技術のアルゴリズム（図2）について述べる．

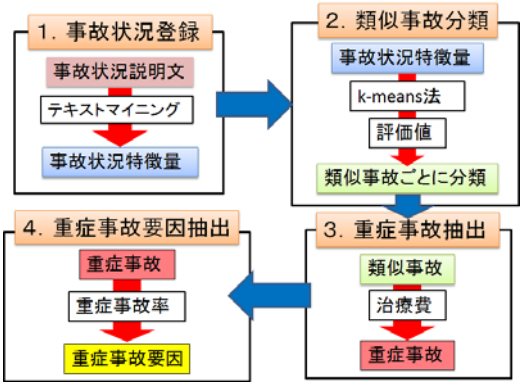

図2 重症度クリフ分析技術のアルゴリズム

事故状況登録では，事故に至るまでの状況が記されている自由記述文に対して，テキストマイニング技術を用いることで，事故の状況を表す事故状況特徴量を登録する．事故状況特徴量とは，事故状況を表すのに重要な単語のことである．ここでは，単語の重要度を，自由記述文から N-gram により単語のつながりを表した行列を作成し，PageRank[2]を用いて評価した．具体的な手順としては，まず，自由記述文の表記ゆれを日本語の概念辞書である日本語 Wordnet[3]を用いて，表現を統一する．次に，形態素解析を行い，ストップワードを削除してから，PageRank を用いて事故状況特徴量を抽出し，登録する．

類似事故分類では，事故状況登録により，登録した事故状況特徴量をもとに類似事故状況ごとに分類する．類似事故の分類は，事故状況特徴量を対象に非階層型クラスタリング手法の 1 つである k-means++法を用いた．クラスタ数 k は，Gap 統計量を用いて最適数を求めた．

重症事故抽出では，分類した類似事故状況下において，重症事故の抽出を行う．ここでは，医療費が高い事故を重傷事故と定義する．類似状況の事故の医療費を，降順に並べてプロットすると，医療費が高い少数事例と医療費が高くない多数事例となり，概ねべき乗則の関係が得られる．重症事故と軽症事故の境目は，統計的変化点検出の手法である偏差の累積和による方法[4]を用いて検出する．

重症事故要因抽出では，類似事故状況下における重症事故と軽症事故を比較することで，何が重症事故の要因かを分析する機能である．重症事故と軽症事故を分類した後，それぞれに出現する単語や特徴を分析する．本研究では，単語や単語の組み合わせなどの条件を指定し，その条件に合致した事故データにおける重症事故が占める割合を重症事故率と定義した．条件として指定する単語や単語の組み合わせなどを変えて，重症事故率を算出することで，重要事故の要因となる特徴を把握可能である．

3. 3 学校環境下の重症事故分析システム

3.1 と 3.2 で述べた技術を統合することで，複数の学校環境下で起きた事故情報を，プライバシを保護したまま統合し，統合した事故データから重症事故を把握したり，その要因を分析することが可能なる．2 つの技術を統合して開発したシステムを図3に示す．

図3 学校環境下の重症事故データ分析システム

このシステムは，データ分析者や教師がユーザとなって利用することを想定している．まず，ユーザが知りたい事故に関する情報（学年，性別，事故の種類，怪我の種類など）を条件として入力することで，その条件に合致する傷害データを複数の学校から取得・統合する．次に，得られた傷害データの事故状況を記述した文章データを対象に重症度クリフ分析

技術を適用することで，指定した条件の事故に関する重症事故を把握可能で，さらにその要因を分析することが可能である．

4. 実データを用いた重症事故分析システムの検証

システムの実用検証を行うために，小学校で発生した事故の傷害データを実際のシステムに提供した．以下に，PSI を用いた傷害データの統合機能の検証と，重症事故分析システムを用いた分析を通した検証について述べる．

4.1 PSI を用いた傷害データの統合機能の検証

PSIを用いた傷害データの統合機能の実用検証を行うために，統合に掛かる処理時間について，対象とする小学校数を変更して検証を行った．具体的には，1 校当たりの傷害データ数は 50～80 件とし，小学校数を 2, 4, 6, 8, 10 校と変え，検索条件は，“性別が男”，“学年が 3 年生”に固定し，同条件で 3 回実行した際の平均の処理時間を求めた．実行環境は，Intel Core i5-5200U，RAM：8GB の Windows8 の PC 上の VirtualBox にセットアップした Ubuntu14．04 である．結果を**図4**に示す．この図は，縦軸が平均処理時間（秒），横軸が小学校数である．この結果から，1機関増えるごとに処理時間が10秒程度増加する傾向があることが分かった．即座に結果を返す検索システムとしての使い方を想定するとユーザの待ち時間があり実用がしにくいが，分析システムとしての利用することを想定し，知りたい条件を指定しておくと，数秒～数時間後に結果を得られる，というシステムであれば，実用可能であると考えられる．その際に，複数の条件を指定でき，順次実行して結果を出力できるシステムにしておくことで，帰宅時に分析を実行しておくと，次の日には複数の分析結果が得られる仕組みなど，実用場面を想定したシステムとすることが重要であることが確認できた．

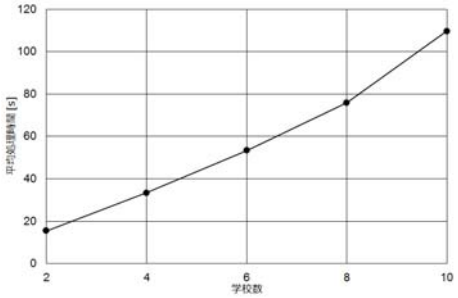

図4 学校数とデータ統合の処理時間の関係

4.2 重症事故分析システムを用いた小学校事故の分析による検証

重症事故分析システムについては，実データに適用することで有用性の検討を行った．小学校で起きた事故（381，493件）を対象に分析を行った事例について述べる．事故データから抽出した事故状況特徴量のうち，頻度が高いものとして，「当たる」，「遊ぶ」，「走る」などが得られた．特に「当たる」と「走る」は同時に現れることが多かったため，「当たる」と「走る」が含まれるクラスタに着目した分析例を紹介する．「当たる」と「走る」が含まれる類似事故状況を対象に重症事故と軽症事故に分類し，重症事故率を算出した．事故が起きた際に行っていた運動を対象に，重症事故率を算出した結果の上位 10 位までを**図5**に示す．

図5より，重症事故率が高い運動は，「筋力トレーニング」，

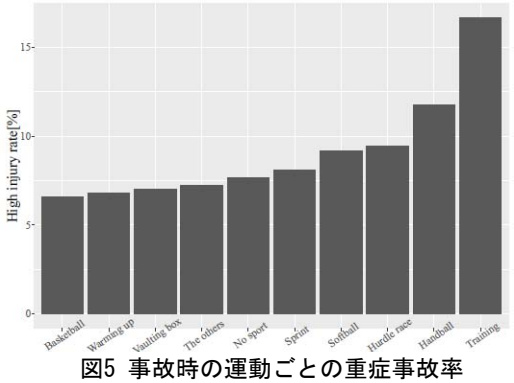

図5 事故時の運動ごとの重症事故率

「ハンドボール」，「障害走（ハードル）」，「ソフトボール」や「短距離走」などがあることが分かる．具体的な事故事例を以下に示す．

1. クラブ活動の時間，ハンドボールをしていた際，ボールを打って 1 塁に走ろうとした時に，運動上に落ちていた小石につまづき倒れ，右足を地面に強くぶつけた．（骨折，18,680 円）
2. 体育の授業中に，校庭で陸上大会のための記録会を行っていたところ，「60m ハードル走」を走っているとき，ハードルに左足をぶつけそのまま転倒した．その際，左足の膝付近を痛めた．（骨折，192,135 円）
3. 体育でかけっこをしていた．となりを走っていた児童がゴール直前に右に曲がり当児童がとなりの児童の足にひっかかり転倒した．そのときに左前頭部を地面で打ち左肘，左ひざには擦過傷ができた．（挫傷・打撲．57,516 円）

具体的な事例を見てみると，ハードルによる傷害は重症度が特に高いことが分かる．ハードルは，練習用のハードルとして，飛び越える横バー部分がゴム製で中間で途切れており，ハードルを飛び損ねても転倒しにくい対策が取られた製品などが作られている．

5. 結論

本稿では，多機関に分散した傷害データをプライバシに配慮したまま共有・統合可能にし，事故状況に着目して重症事故の要因分析を行うためのシステムを提案した．具体的には，PSI と重症度クリフ分析技術を統合したシステムを開発した．実際に，小学校での事故による傷害データに適用することで，システムの有用性を検証した．

今後，条件を指定した重症事故分析だけでなく，日々の軽症事故情報を入力することで，類似した状況下で発生した重症事故を，その要因の特徴も含めて提示するシステムを開発する予定である．

参考文献

1) Atsuko Miyaji, Kazuhisa Nakasho, Shohei Nishida, ``Privacy-Preserving Integration of Medical Data A Practical Multiparty Private Set Intersection″,Journal of Medical Systems,Vol. 41 No. 3, pp. 1-10, (2017)
2) Samer, H., Rada, M. and Carmen B., Random-Walk Term Weighting for Improved Text Classification, International Journal of Semantic Computing, Vol.01 (2007), pp.1-8.
3) Francis, B., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T. and Kanzaki, K., Enhancing the Japanese WordNet, The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP (2009), pp.1-8.
4) Killick, R. and Eckley, I., changepoint: An R Package for Changepoint Analysis, Journal of Statistical Software, Vol.58 (2014), pp.1-19.