Závěrečná zpráva projektu FR CESNET 323/2009

Sběr a zpracování provozních dat pro vyhledávání systémových anomálií

Průběh projektu

V rámci projektu byly zmapovány zdroje provozních dat ve výpočetním prostředí ZČU. Zdroje dat byly roztříděny dle předpokládaného informačního obsahu. Pro další zpracování byly vybrány vhodné zdroje a pořízen dostatečný soubor provozních dat pro další testování. Při sběru dat byly využity i už existující provozní báze dat (centrální logovací servery, datové báze monitorovacích systémů a podobně). Data byla předzpracována a převedena do jednotné platformy, databáze MySQL.

Dalším krokem v projektu byla účast na konferenci IC3K: 2nd Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Valencia Spain. Konference se zúčastnili oba řešitelé. Konference značně obohatila teoretickou základnu řešitelů a mnohé příspěvky inspirovaly k novým pohledům na řešenou tématiku.

Na základě předchozích analýz i dle výstupů konference byly navrženy a otestovány metody selekce a zpracování provozních dat. Preferovány byly ty postupy, které odfiltrují periodické děje v systémech a zvýrazní neperiodické anomálie provozních stavů.

V oblasti vizualizace dat bylo pak vyzkoušeno několik metod zobrazení, s důrazem na globální přehled přes celé výpočetní prostředí. Pro porovnání srozumitelnosti vizualizace provozního stavu byly jako referenční použity systémy MRTG - The Multi Router Traffic Grapher a Nagios, které jsou v ZČU používány po delší dobu v rutinním provozu.

Dosažené cíle

V oblasti sběru dat byl pořízen rozsáhlý vzorek rozmanitých provozních dat za období několika měsíců. Tento vzorek slouží jako testovací množina pro ověřování jednotlivých metod zpracování a vizualizace dat.

Dále byly ověřeny postupy automatizovaného sběru dat z prostředí. Jako velmi výhodné se jeví využití sond systému Nagios, které ve stávajícím prostředí periodicky zjišťují provozní stav většiny systémů. Sondy v posledních verzích už poskytují základní charakteristické provozní údaje, případně je možné generickou sondu požadovaným způsobem snadno upravit, aby žádané údaje poskytovala. Použitím už provozovaných sond Nagios se zároveň nevytváří další duplicitní sběrová síť. Takto provedený systém sběru provozních dat má zároveň k dispozici i funkční monitoring správného běhu sběrových sond, což by bylo při větším počtu sond stejně nezbytné implementovat.

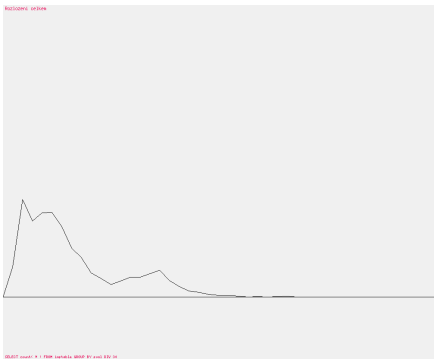
Všechna pořízená data končí v jedné MySQL databázi, nad kterou běží analytický systém. Záznam každého vzorku obsahuje především identifikaci veličiny, systému a subsystému, časový údaj a hodnotu veličiny. Pro zpracování dat byly využity běžné statistické metody, které jsou podporovány funkcemi v jazyce SQL. Důvodem byla právě jednoduchost algoritmů

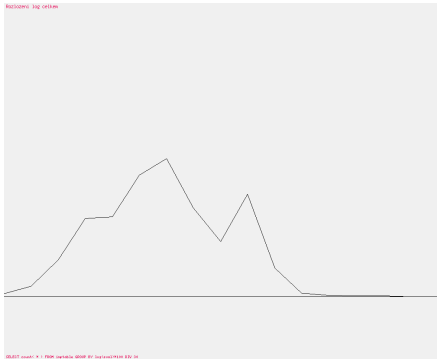
a tím vyšší rychlost při opakovaných výpočtech nad rozsáhlou množinou dat (operativní zobrazování provozních charakteristik).

Základem detekčního algoritmu je porovnání konkrétní hodnoty veličiny s průměrem a směrodatnou odchylkou vzorků. Vzorky ležící uvnitř intervalu odchylky jsou považovány za normální, vzorky ležící mimo za známky anomálie (čím větší vzdálenost od průměru, tím podezřelejší anomálie).

Součástí algoritmu je i filtrace period (denní a týdenní periody). Filtrace spočívá v prostém použití vzorků veličiny pouze z konkrétní hodiny daného dne z celého intervalu vzorků. Testovaný vzorek je tedy porovnáván pouze s hodnotami veličiny naměřenými v tutéž hodinu dne respektive dne v týdnu.

Důležitým předpokladem pro správnou funkci algoritmu je, že testovaná veličina má normální rozdělení. Provozní veličiny obvykle normální rozdělení nemají, protože jsou zdola omezeny, horní mez obvykle nemají (např. počet běžících procesů). Jednoduchým prostředkem dostupným v SQL, jak zlepšit průběh provozní veličiny, je funkce logaritmus. Výsledné rozdělení sice je v rámci možností bližší normálnímu rozdělení, pro toto použití tedy zřejmě dostačující. Následující obrázky ilustrují rozdělení provozní veličiny bez a s použitím logaritmické funkce:





Nedostatkem uvedeného řešení je špatná (resp. žádná) reakce na běžné změny sledovaných veličin, algoritmus se neumí přizpůsobit novým podmínkám. Např. při trvale rostoucí veličině (obsazení disku) je detekována anomálie vždy na začátku a konci intervalu. Stejně tak při zásahu do infrastruktury (např. rozšíření operační paměti serveru), je generována trvale anomálie nového stavu oproti stavu před upgrade.

Vhodným řešením je nasazení algoritmu s klouzavým průměrem a obdobným způsobem počítat i směrodatnou odchylku. Pak se algoritmus přizpůsobí změnám parametrů a vliv starších, méně zajímavých hodnot postupně z výpočtu vymizí.

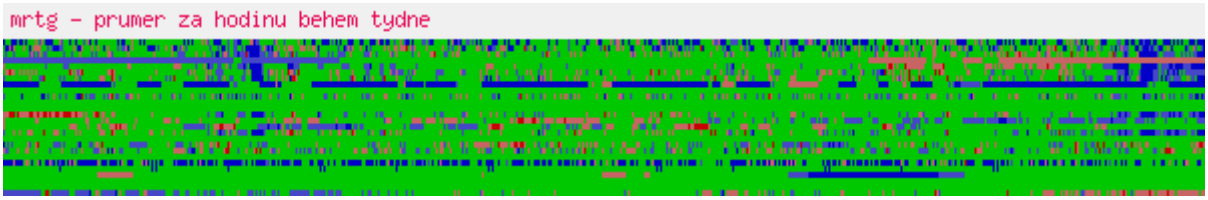
Z vlastností algoritmu nám vyplývají obecnější požadavky na testované veličiny:

- existuje transformace provozního parametru na funkci s normálním rozdělením (alespoň přibližně),
- veličina odráží provozní stav,
- krajní hodnoty veličiny jsou nezdravé (jedna krajní hodnota znamená poruchu, druhá přetížení).

Výsledná funkce pak dává sice často bezrozměrné hodnoty, ale umožňuje vedle sebe hodnotit i zcela nepříbuzné parametry.

Zobrazovací metody byly ověřovány na jednoduché webové aplikaci, generující pro zvolené veličiny a požadovaný časový rozsah výstupní obrazový soubor. Vzorky ležící v pásmu

průměru jsou zelené, vzorky ležící pod pásmem průměru jsou modré (dva stupně), vzorky nad pásmem jsou červené (také dva stupně). Příklad obrazového souboru provozních veličin skupiny poštovních serverů za období dvou měsíců je na následujícím obrázku.



Ve srovnání se stávajícím provozním monitoringem mají použité metody vyšší vypovídající schopnost pro komplexní přehled a zjištění anomálií. Jeden obrazový soubor poskytuje zvolený průřez sledovanými veličinami různých systémů.

Rezervy jsou zatím v možnosti přechodů do detailnějších zobrazení konkrétních provozních parametrů. Testovací program pouze zobrazuje vybranou skupinu dat, další „proklik“ z obrázku na detailnější informace není implementován. Zde se opět nabízí užší integrace zobrazovacích metod se systémem Nagios, který má k dispozici zobrazení časového průběhu jednotlivých veličin. Jednoduchý „proklik“ z obrazového souboru analyzátoru do časového grafu konkrétní veličiny by velmi zrychlil práci při posuzování souvislostí a závažnosti konkrétní anomálie.

Dalším uvažovaným zlepšením detekce anomálií až na grafické úrovni je použití obrazových filtrů z vhodné grafické knihovny přímo na výstupní obrazový soubor (provedení „derivace“ zobrazené veličiny, doostření, zvýraznění hran, odstranění šumu a podobně). Očekávaným výsledkem je optické zvýraznění hlavních anomálií a potlačení méně důležitých odchylek. Tento postup však nebyl zatím prakticky ověřen.

Konkrétní výstupy a další využitelnost

Veškerá dokumentace projektu, výsledky a prezentace je k dispozici na stránkách projektu http://support.zcu.cz/index.php/CIV:Granty/323.

Projekt byl prezentován na semináři CIV, viz http://seminar.civ.zcu.cz.

V rámci projektu byl publikován příspěvek do časopisu DSM (data security management, http://www.dsm.tate.cz), v čísle 2011/2, str. 36, článek „Provozní data a co s nimi“.

Přínosy projektu

Projekt zvýšil informační základnu řešitelského kolektivu v oblasti statistického zpracování dat a vizualizačních metod.

Výstupy tohoto projektu jsou použitelné pro další rozvoj výpočetního prostředí ZČU.

Tisková zpráva

V rámci projektu FR CESNET 323/2009 s názvem „Sběr a zpracování provozních dat pro vyhledávání systémových anomálií“ byla provedena klasifikace provozních dat prostředí. Nasbírané vzorky dat byly zpracovány a na této množině pak byly testovány různé metody zobrazení.

Výstupy projektu jsou k dispozici ostatním členům sdružení CESNET a budou využity při dalším rozvoji výpočetního prostředí Západočeské univerzity.

Příloha

1. Výkaz hospodaření

V Plzni dne 2.2.2011

Jiří Bořík

Výkaz hospodaření

Plánované náklady projektu činily 230 tisíc Kč, skutečné 236 tisíc Kč. Spoluúčast FR CESNET činila dle plánu 150 tisíc Kč. Mírné navýšení spoluúčasti ZČU tvoří cestovní náklady.

Podmínky financování projektu byly dodrženy.