Posudek diplomové práce Jana Waltera „Selected data mining methods and their applicability to the television audience monitoring data in the Czech Republic.

Prof. Dr. Petr Hájek, DrSc, vedoucí.

Cílem práce bylo podat zasvěcený přehled metod „data mining" (těžení z dat) a studovat jejich použití na konkrétní oblast vyhodnocování sledovanosti televize v ČR. Tento cíl si diplomant stanovil sám a měl k dispozici veliká data (viz dále). Pracoval velmi samostatně a s dobrými startovními znalostmi. V práci mj. popisuje moderní obecnou metodiku CRISP pro těžení z dat a srovnává dvě konkrétní metody těžení z dat - nyní populární metodu asociačních pravidel (ve smyslu Agrawala) a „klasickou" českou metodu GUHA. Kromě slovního výkladu diplomant předkládá experiment užití obou metod na data o televizní sledovanosti, která mají asi 3000 objektů a na nich je sledováno asi 1000 otázek (ne všechny jsou zodpovězeny – pracuje se s chybějící informací). Data takových rozměrů zpracoval program pro asociační pravidla uspokojujícím způsobem. Tato metoda připouští jedinou formu pravidel (hypotéz) a jedinou sémantiku (se třemi parametry). Naproti tomu GUHA je teoreticky hluboce propracovaná, umožňuje volbu řady věcí – kvantifikátorů (definujících závislost, souvislost, asociaci), syntaxe hypotéz atd. a proto pracuje pomaleji. Diplomant použil obě metody na celá data (nebo téměř celá data). Počet objektů v tisících není pro metodu GUHA problém, ale tisíce veličin je problém. Proto se diplomant v prováděných bězích metody nedostal k hypotézám s tříčlennými antecedenty ani za více než den běhu. Provedl a analyzoval tři běhy, s různými kvantifikátory. Navrhoval jsem mu, aby využil možností omezit se na několik desítek veličin a rozumně omezit tvar cedentů, to však už nestihl.

Celkově hodnotím práci jako přínosnou pro obecné pojetí těžení z dat a srovnání konkrétních přístupů včetně experimentů na velmi rozsáhlých datech. Pokud je v plánu publikace, doporučuji doplnit současnou verzi o ukázky jemnějšího využití možností metody GUHA. Navrhuji uznat práci jako prácí diplomovou.

V Praze 14.9.2006.